Response patterns in the developing social brain are organized by social and emotion features and disrupted in children diagnosed with autism spectrum disorder

Response Patterns in the Developing Social Brain are Organized by Social and Emotion Features and Disrupted in Children Diagnosed with Autism Spectrum Disorder

Hilary Richardson*, Hyowon Gweon, David Dodell-Feder, Caitlin Malloy, Hannah Pelton, Boris Keil, Nancy Kanwisher, Rebecca Saxe

*Corresponding Author: hilary.richardson@childrens.harvard.edu

Abstract

Adults and children recruit a specific network of brain regions when engaged in "Theory of Mind" (ToM) reasoning. Recently, fMRI studies of adults have used multivariate analyses to provide a deeper characterization of responses in these regions. These analyses characterize representational distinctions *within* the social domain, rather than comparing responses across preferred (social) and non-preferred stimuli. Here, we conducted opportunistic multivariate analyses in two previously collected datasets (Experiment 1: n=20 5-11 year old children and n=37 adults; Experiment 2: n=76 neurotypical and n=29 5-12 year old children diagnosed with Autism Spectrum Disorder (ASD)) in order to characterize the structure of representations in the developing social brain, and in order to discover if this structure is disrupted in ASD. Children listened to stories that described characters' mental states (Mental), non-mentalistic social information (Social), and causal events in the environment (Physical), while undergoing fMRI. We measured the extent to which neural responses in ToM brain regions were organized according to two ToM-relevant models: 1) a condition model, which reflected the experimenter-generated condition labels, and 2) a data-driven emotion model, which organized stimuli according to their emotion content. We additionally constructed two control models based on linguistic and narrative features of the stories. In both experiments, the two ToM-relevant models outperformed the control models. The fit of the condition model increased with age in neurotypical children. Moreover, the fit of the condition model to neural response patterns was reduced in the RTPJ in children diagnosed with ASD. These results provide a first glimpse into the conceptual structure of information in ToM brain regions in childhood, and suggest that there are real, stable features that predict responses in these regions in children. Multivariate analyses are a promising approach for sensitively measuring conceptual and neural developmental change and individual differences in ToM.

1. Introduction

Traditional fMRI analyses compare the average magnitude of response to different experimental conditions in order to discover which brain regions are recruited for a given cognitive task. For example, hundreds of fMRI experiments converge to show that human adults have brain regions that respond preferentially when they consider others' minds – i.e., their beliefs, desires, and emotions (for reviews, see Carrington & Bailey, 2009; Adolphs, 2009). These regions include bilateral temporoparietal junction (TPJ), precuneus (PC), and medial prefrontal cortex (MPFC). While many social tasks recruit the entire network of brain regions, and responses in these regions are correlated even in absence of a task (e.g., Fox et al., 2005; Greicius, Krasnow, Reiss, & Menon, 2003), the response in the RTPJ in particular has been shown to be selective, responding more when participants consider people's mental states relative to other kinds of representations (e.g., photographs; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005), internal states (e.g., pain, hunger, fatigue; Saxe & Powell, 2006; Spunt, Kemmerer, & Adolphs, 2015; Bruneau, Pluta, & Saxe, 2012; Lombardo et al., 2010; Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018), and non-mentalistic social information (e.g., a person's physical appearance or enduring relationships; Saxe & Powell, 2006; Mitchell, Banaji, & Macrae, 2005). As such, the RTPJ has been hypothesized to be particularly important for "Theory of Mind" (ToM) reasoning – our use of an intuitive, structured theory that relates others' actions to their internal, often unobservable, mental states (Gopnik & Wellman, 1992).

However, one limitation of univariate fMRI studies is that, even among neurotypical adults, the magnitude of response isn't particularly sensitive to distinctions *within* the preferred stimulus category. For example, the RTPJ has high responses while processing beliefs regardless of whether they are true or false, or justified or unjustified (Young, Nichols, & Saxe, 2010b; Döhnel et al., 2012). Thus, univariate fMRI analyses are not sensitive to a key question: what aspects of mental states organize and drive responses within ToM brain regions?

Addressing this question may be particularly important for characterizing neural correlates of theory of mind development. As children get older, they increasingly make conceptual distinctions between and based on mental states. For example, while the causal relationship between goals and emotions seems to be understood quite early in development (Repacholi &

Gopnik, 1997; Skerry & Spelke, 2014), children become increasingly aware of causal relations between beliefs and emotions in middle childhood (e.g., after age four years; Harris, Johnson, Hutton, Andrews, & Cooke, 1989; Pons, Harris, & de Rosnay, 2004; Ruffman & Keenan, 1996; Wu & Schulz, 2018). Four- to five-year-old children who correctly report that Little Red Riding Hood (falsely) believes that her grandmother is in the bed nevertheless report that Little Red Riding Hood will feel afraid when she enters her grandmother's home – missing the link between Little Red Riding Hood's false belief and her emotion (Bradmetz & Schneider, 1999). The ability to explicitly distinguish and label the emotions of characters in stories likewise improves throughout middle childhood (Nelson, Widen, & Russell, 2006; Widen, 2016). One intriguing possibility is that as children master new conceptual distinctions between mental states (Gopnik & Wellman, 1992; Koster-Hale & Saxe, 2013), these distinctions also become reflected in neural response patterns in ToM brain regions. If neural response patterns do reflect the conceptual organization of mental states in childhood, they may also provide a window into the nature of theory of mind deficits in neurodevelopmental disorders like autism.

Multivariate approaches have recently been employed to characterize within-category distinctions in neural population responses (Cohen et al., 2017; Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Norman, Polyn, Detre, & Haxby, 2006). While most prevalent in studies of the ventral visual stream (e.g., Haxby et al., 2001, several fMRI studies of adults have used multivariate methods to discover features of mental states that evoke distinct patterns of activity in ToM brain regions (Carter, Bowling, Reeck, & Huettel, 2012; Koster-Hale, Bedny, & Saxe, 2014; Koster-Hale et al., 2017; Koster-Hale, Saxe, Dungan, & Young, 2013; Tamir, Thornton, Contreras, & Mitchell, 2016), and to test hypotheses about the content and structure of representations about other people (Hassabis et al., 2013; Thornton & Mitchell, 2017b; 2017a), and their emotions (Jastorff, Huang, Giese, & Vandenbulcke, 2015; J. Kim et al., 2015; Peelen, Atkinson, & Vuilleumier, 2010; Thornton, Weaverdyck, & Tamir, 2019). For example, Skerry & Saxe (2015) identified three plausible models for the organization of emotion representations, based on prior research: emotion representations could be organized by (1) valence and arousal (the "circumplex" model (Barrett, 2006; Russell, 1980)), (2) six "basic" emotions (Cohen et al., 2017; Du, Tao, & Martinez, 2014; Ekman, 1992), or (3) abstract event appraisals (e.g., "Did a character's emotion involve an event that would or might occur in the

future?"; "Did this situation involve a change in [character's] knowledge or belief about something?" (Ellsworth, 2013; Scherer, 1999)). They found that response patterns in ToM brain regions were best captured by the appraisal model. Neural responses in ToM brain regions to 20 distinct emotions could be classified successfully using this model (Skerry & Saxe, 2015).

To date, though, similar methods have not been applied to capture conceptual change during development. Pediatric fMRI studies typically use univariate measures, like the magnitude and selectivity of the response in ToM brain regions. By age three, ToM brain regions are functionally distinct – they are more correlated with other ToM brain regions than with regions in other functional networks (Richardson et al., 2018; Xiao, Geng, Riggins, Chen, & Redcay, 2019). Responses in ToM brain regions gradually become more selective for reasoning about mental states, relative to non-mentalistic social descriptions (Gweon, Dodell-Feder, Bedny, & Saxe, 2012; Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009) and bodily sensations, like pain, during childhood (Richardson et al., 2018) and adolescence (Richardson, 2019). Increasing sensitivity to category boundaries – i.e., the distinction between preferred and non-preferred stimuli – appears to be one aspect of developmental change in ToM brain regions. Can multivariate approaches capture developmental change or differences in the (within-category) structure of mental state representations?

Initial evidence suggesting that multivariate approaches may be sensitive to differences in the structure of mental state representations comes from studies of adults diagnosed with Autism Spectrum Disorder (ASD), which is a neurodevelopmental disorder characterized by enduring and disproportionate deficits in social and communicative skills (American Psychiatric Association, 2013). While social cognitive deficits are a diagnostic feature of this disorder, and several behavioral studies find evidence for disproportionate deficits on social cognitive tasks in individuals with ASD (e.g., Baron-Cohen, 2000, there is also evidence for variability in the extent of social cognitive deficits, as captured by behavioral tasks (Byrge, Dubois, Tyszka, Adolphs, & Kennedy, 2015; Lombardo et al., 2016; Pierce et al., 2016). And, despite substantial effort, robust, replicable neural correlates of ASD remain elusive (for reviews, see Pelphrey, Shultz, Hudac, & Vander Wyk, 2011; Pelphrey, Adolphs, & Morris, 2004). For example, a recent large-scale study did not find any differences in a range of structural brain measures (e.g.,

cortical thickness, area, and volume, and cerebellar-subcortical measures) between individuals with ASD (n=925, 5-64 years old) and healthy controls (Kaufmann et al., 2019). Similarly, in a relatively large sample of adults, Dufour et al. (2013) did not find any differences in univariate responses in ToM brain regions between neurotypical adults (n=462) and adults diagnosed with ASD (n=31) during a ToM task (Dufour et al., 2013).

A few studies suggest that multivariate analyses may be more sensitive to neural correlates of social deficits in ASD than traditional fMRI analyses. Koster-Hale et al. (2013) measured response patterns in social brain regions in neurotypical (NT) adults and adults with ASD as they read narratives in which someone caused harm to another individual. In NT adults, distinct response patterns were evoked for harm caused accidentally versus intentionally in RTPJ. That is – responses in the RTPJ to stories in which an individual caused harm intentionally looked more similar to responses to other stories that described intentional harm, relative to those that described accidental harm. This distinction was not present in the response pattern in adults diagnosed with ASD (Koster-Hale et al., 2013). Other studies have provided evidence for disrupted response patterns in ASD during attention and mentalizing tasks (Gilbert, Meuwese, Towgood, Frith, & Burgess, 2009), and for a correlation between symptom severity and classification of faces (vs. houses) based on response patterns in the fusiform gyrus (Coutanche, Thompson-Schill, & Schultz, 2011). Note, though, that in another study, multivariate analyses failed to find differences in response patterns between neurotypical adults and adults with ASD during spontaneous processing of emotional facial expressions (Kliemann et al., 2018). Still, multivariate measures of within-category representations could plausibly be more sensitive measures of differences in conceptual representation, both in neurotypical development and in ASD.

Here, we conducted opportunistic analyses of two previously collected pediatric fMRI datasets in order to test whether multivariate approaches are sensitive to the rich within-category structure of mental state representations in children. We used representational dissimilarity matrices (RDMs; Kriegeskorte et al., 2008) in order to measure the pairwise dissimilarity between responses to 24 unique, orally presented child-directed story stimuli originally written to fall into three experimental conditions: Mental (containing explicit descriptions of characters' mental

states: beliefs, desires, emotions), Social (containing descriptions of people and their relationships, but not mental states), and Physical (containing descriptions of causal events in the world, but not people), in brain regions preferentially recruited for mental state reasoning. We then constructed four a priori model RDMs that captured dissimilarity between the stories according to (1) experimenter-generated condition labels (Mental, Social, Physical), (2) emotion content, (3) linguistic features, and (4) narrative features. We chose to use relatively simple ToM-relevant models given the content of the story stimuli, which were written for prior studies and not designed to isolate or vary by ToM-relevant features, and given other methodological limitations of our experiment (at most 24 unique, and complex, story stimuli, with no repetitions, per participant).

Our overarching goal was to test whether multivariate patterns (a proxy for representations) in ToM brain regions change with age, correlate with ToM task performance, and vary by ASD diagnostic status, in childhood. In Experiment 1, we tested whether this experimental paradigm (24 stimuli, each presented only once) allowed for any meaningful measurement of neural population patterns (RDMs). We tested the hypothesis that condition label and emotion models would capture response dissimilarity in ToM brain regions – specifically the right temporoparietal junction – better then the control (linguistic, narrative) models, and tested for developmental differences between children (n=20, 5-12 years old) and adults (n=37). Our a priori region of interest was RTPJ, given the highly selective response profile in adults (e.g., Saxe & Powell, 2006, and prior evidence that developmental change in response selectivity correlates with ToM task performance in childhood (Gweon et al., 2012). We additionally conducted exploratory analyses in the full ToM network (left TPJ, precuneus (PC), and middle medial prefrontal cortex (MMPFC)). We also explored multiple possible models of neural activity patterns: two motivated by a prior fMRI study of emotion representations in adults (Skerry & Saxe, 2015): a circumplex model (Barrett, 2006; Russell, 1980) and an event appraisal model (Ellsworth, 2013; Scherer, 1999), and two that included both condition and emotion features, in a fixed or weighted fashion (Khaligh-Razavi, Henriksson, Kay, & Kriegeskorte, 2017). In Experiment 2, we repeated analyses from Experiment 1 in a large sample of neurotypical children (n=76, 5-12 years old), as well as a smaller sample of children diagnosed with Autism Spectrum Disorder (ASD; n=29, 5-12 years old). We tested for developmental

change in model fits with age and with ToM behavioral score in the neurotypical sample (given the relatively large sample size), and separately tested for disrupted or disordered response patterns in children diagnosed with ASD.

2. Methods

2.1 Preregistration

In addition to reporting information about our participant demographics, tasks, and analyses, we report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

Because this study involved conducting opportunistic analyses of datasets collected between 2009-2012, the study *procedures* were not pre-registered. However, the study procedures in Experiment 2 directly replicated those initially designed for and used in Experiment 1 (Gweon et al., 2012). In order to constrain analysis decisions and to make specific procedures and hypotheses clear, study *analyses* were pre-registered via the Open Science Framework (OSF; https://osf.io/wzd8a; includes preprocessing procedures, region of interest selection and definition, motion exclusion and treatment procedures, calculation of neural response similarity; Asendorpf et al., 2013; Munafò et al., 2017). Analyses were pre-registered specifically for the large sample of neurotypical children (Experiment 2). Exploratory and unplanned analyses are specifically marked as such in the results section, and discrepancies are detailed in the Supplementary Materials.

2.2 Participants

The current study involved conducting opportunistic analyses on previously acquired datasets. As such, sample sizes in the current study were determined based on the number of participants collected for multiple previously conducted studies and pre-registered thresholds for participant exclusion. Sample sizes for the previous studies were not pre-registered, and were determined based on sample size standards at the time (i.e., 2009-2012) and availability of eligible participants.

Experiment 1 was conducted on a previously collected sample of neurotypical adults (n=37, 18-65 years old) and children (n=20 5.1-11.5 year olds, M(SD) age = 8.5(1.8) years, 10 females, 1 left-handed) who completed the story fMRI task. Adults were initially recruited for different studies, and included sighted and right-handed individuals (n=24) and congenitally blind individuals (n=13, 3 left-handed, 2 ambidextrous but right hand preferred). A subset of the sighted adults (n=16) wore a blindfold during the scan because they were recruited as a control sample for studies on plasticity in the visual cortex in individuals who are blind. Results of univariate analyses of the children and adults in Experiment 1 have been previously published (Gweon et al., 2012; Bedny, Richardson, & Saxe, 2015).

Experiment 2 participants were 76 neurotypical children (NT; 16 females, M(SD) age = 8.6(2.0) years, range: 5.3-12.6 years, handedness: 4 left-handed, 1 ambidextrous, 10 NA), and 29 children diagnosed with Autism Spectrum Disorder (ASD; 4 females, M(SD) age = 9.5(1.7), range: 5.6-12.9 years, handedness: 3 left-handed, 1 ambidextrous, 4 NA). Criteria for ASD status included both a clinical diagnosis of autism, Asperger's, or PDD-NOS (DSM-IV) by a specialist in neurodevelopmental disorders, and a classification of 'autism' or 'autism spectrum disorder' on the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000) conducted by a research-reliable administrator. All children who participated in Experiment 2 had a standardized IQ score > 80, as measured by the non-verbal Kaufman Brief Intelligence Test (KBIT-2; Kaufman, 1997). An additional 7 neurotypical children and 25 children diagnosed with ASD were recruited but excluded from analyses due to not completing at least two functional runs of the fMRI experiment (n=3 NT, n=7 ASD), excessive motion during the scan (n=4 NT, n=17 ASD), or incidental findings in the structural MRI data (n=1 ASD). Experiment 2 data have not previously been published.

Neurotypical children were recruited using local parenting listservs, promotional activities, and flyers at libraries and museums. Children diagnosed with ASD were recruited using existing clinical databases (Simons Simplex Collection, SFARI, Autism Consortium). All participants were recruited from the New England area, were native speakers of English, and had no known other neurological or cognitive disabilities. All children gave written assent, and their parents

gave written informed consent, in accordance with the requirements of the Committee on the Usage of Humans as Experimental Subjects at MIT.

2.3 Behavioral Battery

Child participants completed a custom-made theory of mind behavioral battery. This task assessed participants' ability to make predictions and provide explanations about the beliefs, desires, actions, and emotions of various characters in a story. The ToM concepts included in this booklet were largely drawn from work describing the successive ToM achievements in early childhood (Wellman & Liu, 2004), with the addition of questions involving reasoning about moral blameworthiness. ToM booklet stimuli are available via OSF ("Booklet 1" on https://osf.io/cbw6f/), and have been described in a prior study (Gweon et al., 2012). The ToM behavioral battery was video-recorded and coded offline by an undergraduate research assistant; the summary score of this measure is calculated as the proportion of questions answered correctly.

Experiment 2 participants additionally completed a measure of non-verbal IQ (KBIT-II Matrices task; Kaufman, 1997). Age-standardized IQ scores were calculated based on the provided protocol.

2.4 FMRI Experiment

Participants listened to English stories involving characters and their mental states (Mental condition), characters and their appearance or social relationships (Social condition), or descriptions of physical objects and events in the world (Physical condition). This experimental paradigm was designed for use with children, and has been described in prior publications (Bedny et al., 2015; Gweon et al., 2012). The story stimuli are publicly available (https://osf.io/cbw6f/). Each story was read by one of three female speakers in child-directed prosody. Stories were matched across condition for number of words (M=52.5 words), number of sentences (4.7), length (20s), and Flesch Reading Ease Level (M=85.7). Story properties were quantified using CohMetrix (http://tool.cohmetrix.com/; McNamara, Louwerse, Cai, & Graesser, 2013).

After each story (20s), participants were asked, “Does this come next?” (1.5s) They then heard a clip containing the story ending or the ending of an unrelated story (3s), followed by an 6.5s pause during which they responded to the prompt by pushing one of two buttons (“Yes” or “No”). This was followed by an encouragement clip: “Way to go!” for correct responses, or “Let’s try another!” for incorrect responses (5s). Half of the presented stories were followed by the correct ending (“Yes” response). Incorrect endings were drawn randomly from all other English story conditions. Analyses of the fMRI data including only hemodynamic responses during the initial 20s story.

Stimuli were presented in Matlab 7.6 (Exp. 1) or Matlab 2010a (Exp. 2) running on an Apple MacBook Pro. Participants heard 24 stories (8 per condition) across four 6.6-minute runs. Participants also heard 8 clips of instrumental music and 8 stories read in a foreign language; these conditions were excluded from the present analyses. Each run included ten 36s blocks (2 per condition), as well as 12s rest at the beginning, halfway point, and end. The order of conditions in each run was palindromic (e.g., [rest] A B C D E [rest] E D C B A [rest]) and counterbalanced across runs. Stories were counterbalanced across runs and participants. A colorful swirl image was presented visually during the stories, as well as during the rest period. During the prompt, story ending, and response portion of the experiment, an image of a check (left) and an “X” (right) was displayed to encourage participants to answer the question, and to remind them which buttons corresponded to “yes” and “no” answers. Participants were introduced to the task and completed five practice trials prior to the scan.

During the scan, child participants were monitored by an experimenter in the control room and a second experimenter who stood next to the scanner bore. If the participant moved noticeably during the scan, this experimenter would place her hand on the child’s leg, as a reminder to stay still.

Attention to the stories was verified by measuring accuracy (proportion of questions answered correctly) on the “Does this come next?” task. Accuracy was calculated using trials from included functional runs and conditions only (trials from runs excluded due to excessive motion were not analyzed). Overall, participants performed well on this task, indicating good attention to

the stories (M(SE) Accuracy **Exp. 1:** Children: .92(.02), Adults: .99(.004); **Exp. 2:** NT Children: .88(.02), ASD Children: .90(.03)). In Experiment 1, adults were more accurate than children (effect of age: b=-.76, t=-4.2, p=.0001); there were no effects of condition (bs<|.14|, ts<|.1|, ps>.3) and the condition-by-age group interaction was not significant. In Experiment 2, there was a significant positive effect of age on performance among neurotypical children (n=76; b=.39, t=4.7, $p=1.2x10^{-5}$), no effect of condition (bs<.13, ts<1.1, ps>.3), and no significant condition-by-age interactions. In the full sample, there was no main effect of group (NT vs. ASD: b=-.11, t=-.59, p=.56) or condition (bs<|.09|, ts<|1|, ps>.3), and the group-by-condition interactions were not significant.

2.5 FMRI Data Acquisition

Prior to the fMRI scan, child participants watched a movie of their choice in a mock scanner while practicing lying still on their back and listening to a recording of scanner sounds for 10-15 minutes. If participants moved during the mock scan, their movie paused for three seconds, reminding and training them to stay still. Mock scanning often reduces participant motion, especially among pediatric samples (de Bie et al., 2010).

Whole-brain structural and functional MRI data were acquired on a 3-Tesla Siemens Tim Trio scanner located at the Athinoula A. Martinos Imaging Center at MIT. Experiment 1 participants used the standard Siemen's 12-channel head coil. Experiment 2 participants used one of two custom 32-channel phased-array head coils made for younger (n=18 NT, n=4 ASD) or older (n=37 NT, n=20 ASD) children (Keil et al., 2011) or the standard Siemens 32-channel head coil (n=19 NT, n=4 ASD; coil information not available for n=2 NT, n=1 ASD). T1-weighted structural images were collected in 128 (Exp. 1) or 176 (Exp. 2) interleaved sagittal slices with 1.33mm (Exp. 1) or 1mm isotropic voxels (Exp. 2; GRAPPA parallel imaging, acceleration factor of 3; adult coil: FOV: 256mm; pediatric coils: FOV: 192mm). Functional data were collected with a gradient-echo EPI sequence sensitive to Blood Oxygen Level Dependent (BOLD) contrast in 3x3x4mm (Exp. 1) or 3mm isotropic (Exp. 2) voxels in 30 (Exp. 1) or 32 (Exp. 2) interleaved near-axial slices aligned with the anterior/posterior commissure, and covering the whole brain (EPI factor: 64; TR: 2s, TE: 30ms, flip angle: 90°). Prospective acquisition correction was used to adjust the positions of the gradients based on the participant's

head motion one TR back (Thesen, Heid, Mueller, & Schad, 2000). 198 volumes were acquired in each run, and functional data were acquired across four runs. Four dummy scans were collected to allow for steady-state magnetization.

2.6 FMRI Data Analysis

2.6.1 Preprocessing

All preprocessing decisions, including procedures for excluding timepoints and participants due to motion, were pre-registered. FMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software written in Matlab. Functional images were registered to the first image of the first run; that image was registered to each participant's anatomical scan, and each participant's anatomical scan was normalized to a common brain space (Montreal Neurological Institute (MNI) template). All data were smoothed using a Gaussian filter (5mm kernel).

Motion artifact timepoints were identified using the ART toolbox (https://www.nitrc.org/projects/artifact_detect/; Whitfield-Gabrieli, Nieto-Castanon, & Ghosh, 2011) as timepoints for which there was 1) more than 2mm of motion in any direction relative to the previous timepoint or 2) a fluctuation in global signal that exceeded a threshold of three standard deviations from the mean global signal. Runs were excluded from analyses if one-third or more of the timepoints collected were identified as motion artifact timepoints, and participants were excluded from all analyses if they had fewer than two runs of usable data (Exp. 2: n=4 NT, n=17 ASD). In both experiments, the total number of motion artifact timepoints was highly correlated with mean translation (henceforth, "motion") – i.e., the average amount of motion (mm) in x, y, z directions between each image, including images identified as motion artifacts (**Exp. 1:** r=.58, p=2.8x10^{-6}, **Exp. 2:** r=.75, p<2.2x10^{-16}).

We tested whether motion differed by variables of interest in each experiment. In Experiment 1, motion did not differ by age group (M(SD) Children = .12(.06), Adult = .10(.04), Cohen's d=-.39 (small)). In Experiment 2, motion was uncorrelated with age and ToM among neurotypical children (age: r(74)=-.14; ToM: r(73)=.13), and did not differ between NT children and children with ASD (M(SD) NT = .15(.07), ASD = .14(.07), Cohen's d=-.04 (negligible)). See

Supplementary Figure 1 for a visualization of motion by experiment and sample, and Supplementary Table 1 for amount of motion per participant in Experiment 2. Despite not differing by the variables of interest (age, group), motion was included as a covariate in all linear regression models that tested for between-subject and between-group differences in model fits. Note that any within-subject or within-group comparison of model fits cannot be driven by motion, as the models are fit to the exact same neural data.

A final strategy for combating potential contamination of motion artifact was to generate five aCompCor regressors (Behzadi, Restom, Liau, & Liu, 2007) from individual white matter masks (eroded by two voxels, to avoid partial voluming), and to include these regressors in the models that estimated betas per item and condition (see section 2.6.2, below). FMRI data were high-pass filtered (threshold: 1 cycle/128 seconds) in order to remove low-frequency fluctuations in the fMRI signal, after interpolating over artifact timepoints (Carp, 2013; Hallquist, Hwang, & Luna, 2013).

2.6.2 Models for Multivariate and Univariate Analyses

We used two general-linear models to analyze BOLD activity of each participant as a function of (1) item, for multivariate analyses, and (2) condition, for supplementary univariate analyses. Data were modeled in SPM8 using a standard hemodynamic response function (HRF). Boxcar regressors for each (1) item or (2) condition were convolved with the standard HRF, and nuisance covariates were included for run effects, motion artifact timepoints, and signals of no interest (five aCompCor regressors; Behzadi et al., 2007).

2.6.3 Defining Individual Regions of Interest

Given the small amount of data per participant, and the high dimensionality of fMRI data, feature selection was used to identify voxels likely to contain relevant information (De Martino et al., 2008; Pereira, Mitchell, & Botvinick, 2009). Within each ROI search space, we defined individual ROIs as the 80 voxels with the highest T-value to an all stories (MSP) > rest contrast, within 10mm sphere hypothesis spaces. This univariate selection procedure helps to eliminate high-variance, noisy voxels (Mitchell et al., 2004), eliminates differences in the number of voxels across ROIs and participants, and is orthogonal to subsequent multivariate analyses. The

choice of 80 voxels was pre-registered (https://osf.io/wzd8a) and based on prior work that conducted multivariate analyses to characterize responses in ToM brain regions (Kliemann et al., 2018; Skerry & Saxe, 2014). Hypothesis spaces were 10mm spheres drawn around peak coordinates for 462 neurotypical adults to a theory of mind localizer task, as described in Dufour et al., 2013. These hypothesis spaces are publicly available for download (http://saxelab.mit.edu/use-our-theory-mind-group-maps).

2.6.4 Neural Representational Dissimilarity Matrices

We calculated a neural representational dissimilarity matrix (RDM) for the story stimuli (n=24), per subject and region of interest (bilateral TPJ, MMPFC, and PC). To do so, we extracted T-values from each voxel within each ROI to each item, and calculated the Euclidean distance (square root of distance*distance) between each pair of stories, across all voxels. Extracting T-values (rather than beta estimates) increases classification performance of linear support vector machines (Misaki, Kim, Bandettini, & Kriegeskorte, 2010) and effectively noise-normalizes the neural RDM (Walther et al., 2016). We normalized each subject's RDM by subtracting the minimum and dividing by the range of values, per subject. See Supplementary Figure 2 for a visualization of average neural RDMs per experiment, sample, and ROI.

The noise ceiling was calculated per region of interest and experiment by creating an average neural RDM across all but one participant (per experiment), calculating the Kendall tau correlation between this average RDM and the neural RDM from the left out participant, and iterating across participants. Regions in which the noise ceiling was significantly above chance in Experiment 2 (n=105) were included in statistical analyses; this resulted in excluding DMPFC and VMPFC (see Supplementary Figure 3 for noise ceilings per ToM ROI, experiment, and sample). Importantly, the noise ceiling did not differ by variables of interest in either experiment: in Experiment 1, the noise ceiling did not differ by age group (**RTPJ:** effect of age group: b=.23, t=.80, p=.43, effect of motion: b=.03, t=.25, p=.81; **all analyzed ROIs:** effect of age group: b=.04, t=.30, p=.76, effect of ROI (LTPJ): b=-.05, t=-.29, p=.77, effect of ROI (PC): b=.35, t=2.0, p=.049, effect of ROI (MMPFC): b=-.54, t=-3.1, p=.003, effect of motion: b=-.07, t=-1.1, p=.30, no group-by-ROI interactions). In Experiment 2, the noise ceiling did not differ by age or ToM among neurotypical children (**RTPJ:** effect of age: b=.03, t=.29, p=.78, effect of motion:

b=-.28, t=-2.5, p=.01, no age-by-motion interaction; effect of ToM: b=.06, t=.56, p=.58, effect of motion: b=-.32, t=-2.8, p=.007, no ToM-by-motion interaction; **all analyzed ROIs:** effect of age: b=.08, t=1.2, p=.25, effects of ROIs: bs<|.16|, ts<|1.2|, ps>.2, effect of motion: b=-.19, t=-2.6, p=.01, no significant interactions; effect of ToM: b=.09, t=1.3, p=.21, effects of ROIs: bs<|.19|, ts<|1.4|, ps>.18, effect of motion: b=-.35, t=-3.0, p=.003, ROI (MMPFC)-by-motion interaction: b=.32, t=2.3, p=.02, all other interactions were non-significant). The noise ceiling did not differ between neurotypical children and children with ASD (**RTPJ:** effect of group: b=-.01, t=-.06, p=.95, effect of motion: b=-.22, t=-2.2, p=.03, no group-by-motion interactions; **all analyzed ROIs:** effect of group: b=.11, t=.80, p=.42, effects of ROIs: bs<|.14|, ts<|1.2|, ps>.2, effect of motion: b=-.19, t=-3.1, p=.002, no significant interactions). This suggests that age, ToM, and group effects on model fits are unlikely to be driven by differences in the reliability of the neural RDMs.

2.6.5 Departures from Preregistered fMRI Analyses

FMRI analyses were pre-registered on the Open Science Framework (OSF; https://osf.io/wzd8a). Based on prior studies relating neural development to theory of mind in children (Gweon et al., 2012; Sabbagh, Bowman, Evraire, & Ito, 2009), we planned to conduct our primary fMRI analyses on right temporoparietal junction (RTPJ) and dorsomedial prefrontal cortex (DMPFC), and to conduct exploratory analyses of responses in other ToM brain regions. However, upon calculating the noise ceiling for each ToM brain region and sample, we found that we could not reliably estimate model fits to data extracted from DMPFC or VMPFC (Supplementary Figure 3). Subsequent statistical analyses of the model fits for these regions were not conducted; though see Supplementary Figures 6, 8, and 10 for visualizations of model fits in these regions. Additional departures from the pre-registered analyses are described in the Supplemental Materials.

2.7 Model Representational Dissimilarity Matrices (RDMs)

2.7.1 Planned Model RDMs

To test whether ToM-relevant features capture the pattern of activity (and developmental change) in ToM brain regions, we created a two model RDMs: (1) a condition label RDM, which reflected binary Mental, Social, and Physical condition labels, and (2) an emotion feature RDM,

which used adult Amazon's Mechanical Turk ratings of seven emotions: embarrassed, joyful, surprised, angry, disappointed, afraid, and hopeful. We compared the fit of these RDMs to two control models: (1) a linguistic features RDM, created using Coh-Metrix (http://tool.cohmetrix.com/; McNamara et al., 2013) ratings of three linguistic features: (1) word count, (2) number of words before the main verb (a measure of syntactic simplicity, and often correlated with working memory demands), and (3) concreteness (a measure of semantic cohesion/ease of understanding), and (2) a narrative features RDM, created using adult MTurk ratings of three narrative features: engagingness, ease of imagination or visualization, and amount of magic/fantasy. See Figure 1 for a visualization of the models.

Emotion and narrative feature ratings were acquired via Amazon's Mechanical Turk. Adults (n=25 unique workers) read a single story per "HIT," and were asked to use a Likert scale (1-7) to indicate "How much does someone in the story feel [emotion]?" A single HIT asked a worker to rate all seven emotions per story. The order of emotions was randomized across the 24 stories/HITs. Workers were then asked to rate (in order): (1) "How engaging was the story?", (2) "How easy was it to imagine or visualize the story?", and (3) "To what extent does this story involve magic and/or fantasy?". Finally, workers were prompted to type the second-to-last word of the story into a blank box; ratings were analyzed if workers passed this quality control item. Workers were allowed to provide ratings for as many of the (24) stories as they wanted, and completed an average of 15 HITs (standard deviation = 9.4). Each story was rated by 14-16 unique workers (M(SD)=15.6(.58)).

Adult emotion and narrative feature ratings were used such that the model RDMs theoretically reflected the mature representational dissimilarity space of the story stimuli. Therefore, we expected that developmental change among children would manifest as increases in the fit of the neural RDMs to the condition and emotion models, with age.

Model RDMs, like neural RDMs, were generated by calculating the Euclidean distance between each story across all ratings/features, and normalizing this distance. Our a priori RDMs were at most moderately positively correlated (**Cond-Emo:** r=.16, **Cond-Ling:** r=.02, **Cond-Narr:** r=-

.03, **Emo-Ling:** *r*=.13, **Emo-Narr:** *r*=.12, **Ling-Narr:** *r*=.09; see Supplementary Figure 4 for correlations between all planned and exploratory model RDMs).

2.6.2 Exploratory Model RDMs

We additionally constructed four exploratory model RDMs after conducting initial analyses using the a priori models in Experiment 1, in order to determine if we could find a model that better captured response patterns in ToM brain regions, and that was more sensitive to developmental change with age (Figure 1). The four exploratory models included two based on a prior fMRI study in adults (Skerry & Saxe, 2015): (1) a circumplex model based on valence and arousal features (Barrett, 2006; Russell, 1980), and (2) an emotion appraisal model based on 38 emotion appraisal features (Ellsworth, 2013; Scherer, 1999). Valence, arousal, and emotion appraisal features were acquired via an independent Amazon's Mechanical Turk study. Adults (n=81 unique workers) read a single story per "HIT", and used a Likert scale (1-7) to rate 38 emotion appraisal statements, as well as valence and arousal. Workers could provide ratings for as many of the stories as they wanted, and rated 7 stories on average (standard deviation = 8.0). Each story was rated by 21-25 unique workers (M(SD)=24.4(.1.1)). Because the fMRI stimuli used here were not developed with these features in mind, and the features are quite specific (e.g., "Did this story involve events consistent with a character's personality or self-concept?"), we calculated the split-half reliability of ratings across 100 random split-half iterations, per feature, and only included features that had good reliability (mean $r>.8$; 34/38 emotion appraisal features and 2/2 circumplex features) in the emotion appraisal and circumplex RDMs.

The other two exploratory models were constructed based on the a priori condition and emotion models. Given that these two models explained some variance in neural responses and outperformed the control models in Experiment 1, and were only moderately positively correlated (*r*=.16), we hypothesized that an optimal combination of the features from each might explain the most variance in neural responses. We constructed an exploratory RDM using both emotion (7) and condition (3) features (a "Emotion-Condition" (EC) model), and a weighted emotion-condition (WEC) model (Khaligh-Razavi et al., 2017), using a non-negative least squares algorithm (Jozwik, Kriegeskorte, & Mur, 2016) to find single-dimension RDM weights that best predicted the average neural RDM, per ROI, in the Experiment 1 sample. Weights were

estimated iteratively on 22/24 stimuli, predicting the fit on the left out 2 stimuli. See Supplementary Figure 5 for a visualization of features.

2.7 Statistical Analyses

2.7.1 Individual Region of Interest Analyses

First, we compared each model RDM's fit to the RDM of the a priori ROI (RTPJ) to chance (0), using one-tailed Wilcoxon signed rank tests. We subsequently directly compared the fit of different models, first in the RTPJ (using Wilcoxon signed rank tests), and then in mixed effects linear regressions that included data from all ROIs (R/LTPJ, PC, MMPFC) and tested for a main effect of and interaction by ROI. Non-significant interaction terms were removed from regressions. These regressions included subject ID as a random effect, in order to account for non-independence of data across ROIs.

In Experiment 1, we tested for age-related change by directly comparing model fits across children (n=20) and adults (n=37). In Experiment 2, we (i) conducted sensitive tests for age- and ToM-related change in model fits in a large sample of neurotypical children (n=76; using continuous variables for age and ToM), and (ii) tested for a group difference in model fits between neurotypical children and children diagnosed with Autism Spectrum Disorder (n=29).

Figure 1

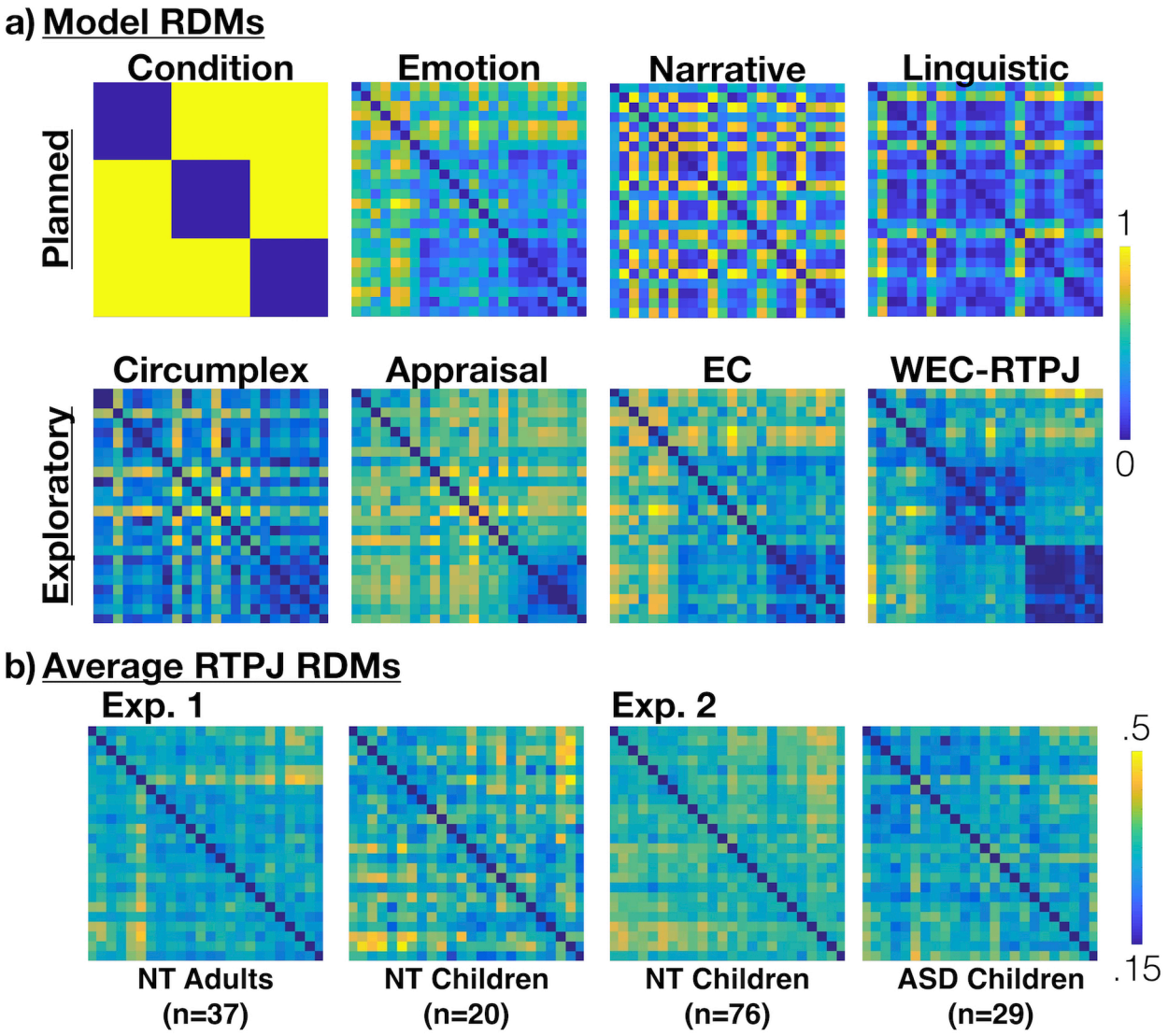


Figure 1. Model and RTPJ Representational Dissimilarity Matrices (RDMs). a) Model RDMs. We measured the extent to which four planned models (top row) and four exploratory models (bottom row) captured dissimilarity in neural response patterns to 24 unique, orally presented story stimuli. The condition model reflected the conditions that the stimuli were originally written to fall into (Mental, Social, Physical). Stories were rated on emotion, circumplex, and appraisal features in studies posted on Amazon's Mechanical Turk, and linguistic features were measured using CohMetrix. The "EC" model refers to a model that uses both emotion and condition features to represent dissimilarity across stories. The WEC (Weighted Emotion-Condition) used a non-negative least squares algorithm to weight each emotion and condition feature, in order to best predict the average neural RDM in the Experiment 1 sample (per ROI; the RTPJ-derived model is shown here). The dissimilarity scale ranges from 0 (similar) to 1 (dissimilar). b) Average RTPJ RDMs, per experiment and sample.

2.7.2 Exploratory Searchlight Analysis

We conducted a searchlight analysis in the combined neurotypical child sample (n=96, across Exp. 1 and Exp. 2) to complement the ROI analyses, and to ensure that unpredicted effects did

not go unnoticed. We defined a 9mm radius sphere surrounding every voxel within a grey matter mask (125575 spheres total), and identified the 80 voxels with the highest t-values to the all stories (Mental/Social/Physical) > Rest contrast within each sphere. We extracted T-values from these 80 voxels to each item, and calculated the Euclidean distance between each pair of stories, across voxels, and normalized the resulting RDM. We then calculated the Kendall tau correlation between each neural RDM and the condition and emotion model RDMs. We created an image of the z-scored Kendall tau correlation values assigned to each voxel per subject, and conducted a whole-brain random effects analysis on the resulting images in order to visualize voxels that show activity correlated with each model. Analyses were corrected for multiple comparisons by estimating the false-positive rate via 5,000 Monte Carlo permutations using the SnPM toolbox for SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/), at $p<.05$.

2.8 Data and Resource Availability

Because these data were collected up to ten years ago, and prior to the normalization of data sharing, the conditions of our ethics approval did not include public archiving of individual raw MRI or behavioral data. That is, participants and parents of participants did not agree to their data being shared publicly. Individuals seeking access to any raw data should contact the last author (Rebecca Saxe; saxe@mit.edu). Access will be granted to individuals who complete a formal data usage agreement through the Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT. Summary data, analysis code, and stimuli are publicly available for download (https://osf.io/cbw6f/).

3. Results

3.1 Experiment 1

In Experiment 1, we first tested whether the experimental paradigm allowed for any meaningful measurement of neural patterns expressed in representational dissimilarity matrices, given that these were opportunistic re-analyses of existing data collected on an experiment that was not designed with multivariate analyses in mind (i.e., relatively few stimuli, with no repetitions). Despite these limitations, we found evidence that ToM-relevant features organized neural response patterns in ToM brain regions. In the RTPJ, patterns of neural activity were correlated with both the condition and emotion models, significantly better than chance (Wilcoxon Signed-

rank tests, chance = 0; **Cond:** M(SE)=.03(.01), W=1130, p=.003; **Emo:** .05(.01), W=1298, p=.0009); the linguistic and narrative control models did not differ significantly from chance (**Ling:** .003(.01), W=739, p=.76; **Narr:** -.01(.01), W=560, p=.98). The emotion and condition models did not differ in their fit to the RTPJ RDM (W=648, p=.22; two-tailed). The condition and emotion models each performed significantly better than either control model (**Cond vs. Ling:** W=1120, p=.004; **Cond vs. Narr:** W=1225, p=.0003; **Emo vs. Ling:** W=1241, p=.0005; **Emo vs. Narr:** W=1319, $p=4.6x10^{-5}$). The same pattern of results was found across all ToM ROIs (mixed effects linear regressions, see Table 1 for full statistics). Across all ROIs, the two ToM-relevant models did not differ in their fit to the neural RDMs (Table 1).

We then tested for group differences in the fit of the two ToM-relevant feature models, based on age group (adult vs. child). In the RTPJ, there were no differences in the fit of either model (**Cond:** M(SE) Adult=.03(.01), Child=.03(.02); **Emo:** Adult=.05(.01), Child =.04(.02); effect of age group: bs<.04, ts<.14, ps>.8; controlling for motion). Similarly, there was no effect of age group on the fit of these models across all ROIs (effects of age group: bs<|.21|, ts<|1.2|, ps>.2, effects of ROIs: bs<|.25|, ts<|1.4|, ps>.17). However, there was an age-by-ROI interaction such that the condition model fit was higher in children, relative to adults, in MMPFC, relative to RTPJ (b=.70, t=2.1, p=.04). See Figure 2.

Figure 2

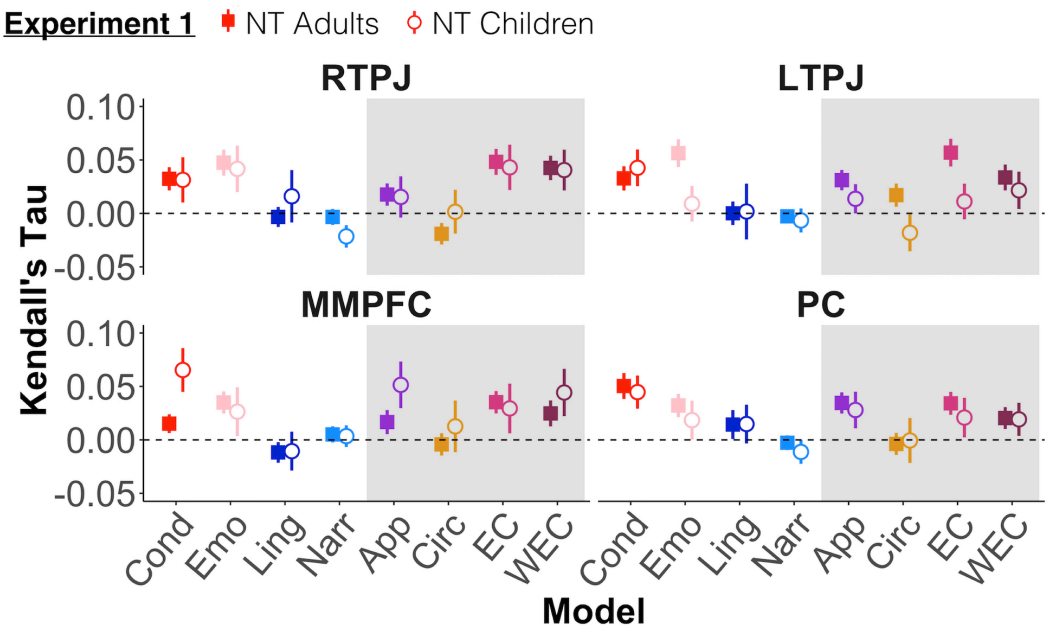


Figure 2. Model Fits in Experiment 1. Plots show the mean Kendall tau correlation (y-axis) between each model (x-axis) and individual neural RDMs, per ROI (RTPJ, LTPJ, PC, MMPFC). Filled squares represent means calculated from adults (n=37); open circles represent means calculated from children (n=20). Lines surrounding mean values indicate standard error from the mean. ToM-relevant (Condition, Emotion) models are shown in red/pink; control (Linguistic, Narrative) models are shown in blues. The shaded area indicates exploratory models, which included a model based on abstract appraisal features (App, purple), a circumplex model based on valence and arousal (Circ, yellow), and models that included both emotion and condition features (EC, hot pink; W (weighted) EC, maroon)).

Table 1

Condition vs. Linguistic	Experiment 1	Experiment 2 (NT)	Experiment 2 (ASD)
Model (Ling)	**b=-.46, t=-5.3, p= 1.8x10^{-7}**	**b=-.40, t=-5.3, p=1.8x10^{-7}**	b=.07, t=.31, p=.76
ROI (LTPJ)	b=.01, t=.09, p=.93	**b=-.23, t=-2.2, p=.03**	**b=.77, t=3.2, p=.002**
ROI (MMPFC)	b=-.09, t=-.73, p=.47	b=-.10, t=-.91, p=.36	b=.43, t=1.8, p=.08
ROI (PC)	b=.18, t=1.5, p=.14	b=.09, t=.88, p=.38	b=.22, t=.93, p=.35
Model (Ling) x ROI (LTPJ)			**b=-1.2, t=-3.5, p=.0005**
Model (Ling) x ROI (MMPFC)			**b=-.78, t=-2.3, p=.02**
Model (Ling) x ROI (PC)			b=-.29, t=-.86, p=.39
Condition vs. Narrative	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Narr)	**b=-.65, t=-7.5, p=3.1x10^{-13}**	**b=-.75, t=-5.1, p=5.5x10^{-7}**	b=.04, t=.16, p=.87
ROI (LTPJ)	b=.08, t=.63, p=.53	b=-.17, t=-1.1, p=.27	**b=.89, t=3.9, p=.0001**
ROI (MMPFC)	b=.12, t=.97, p=.33	b=-.25, t=-1.7, p=.10	**b=.50, t=2.2, p=.03**
ROI (PC)	b=.16, t=1.3, p=.19	b=-.05, t=-.32, p=.75	b=.26, t=1.1, p=.26
Model (Narr) x ROI (LTPJ)		b=.28, t=1.3, p=.18	**b=-1.1, t=-3.6, p=.0004**
Model (Narr) x ROI (MMPFC)		**b=.60, t=2.8, p=.005**	b=-.58, t=-1.8, p=.07
Model (Narr) x ROI (PC)		b=.22, t=1.1, p=.29	b=-.32, t=-.99, p=.33
Emotion vs. Linguistic	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Ling)	**b=-.43, t=-5.0, p=7.9x10^{-7}**	**b=-.19, t=-2.6, p=.009**	b=-.16, t=-1.3, p=.21
ROI (LTPJ)	b=-.05, t=-.44, p=.66	b=-.15, t=-1.4, p=.15	b=-.19, t=-1.1, p=.27
ROI (MMPFC)	b=-.18, t=-1.5, p=.14	b=-.02, t=-.18, p=.86	b=-.19, t=-1.1, p=.28
ROI (PC)	b=-.04, t=-.36, p=.72	b=.10, t=1.0, p=.32	b=-.11, t=-.65, p=.51
Emotion vs. Narrative	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Narr)	**b=-.60, t=-7.0, p=9.7x10^{-12}**	**b=-.22, t=-3.1, p=.002**	b=-.09, t=-.73, p=.47
ROI (LTPJ)	b=-.0006, t=-.005, p=.996	b=.04, t=.42, p=.68	b=-.11, t=-.61, p=.54
ROI (MMPFC)	b=.006, t=.05, p=.96	b=.12, t=1.2, p=.24	b=-.08, t=-.43, p=.67
ROI (PC)	b=-.11, t=-.88, p=.38	b=.08, t=.78, p=.44	b=-.12, t=-.67, p=.51
Condition vs. Emotion	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Emo)	b=-.02, t=-.21, p=.83	**b=-.18, t=-2.3, p=.02**	b=.09, t=.41, p=.68
ROI (LTPJ)	b=-.01, t=-.11, p=.91	b=-.07, t=-.68, p=.50	**b=.74, t=3.2, p=.002**
ROI (MMPFC)	b=-.09, t=-.71, p=.48	b=-.13, t=-1.2, p=.24	b=.41, t=1.8, p=.07
ROI (PC)	b=-.02, t=-.12, p=.90	b=-.01, t=-.13, p=.90	b=.21, t=.93, p=.35
Model (Emo) x ROI (LTPJ)			**b=-.73, t=-2.3, p=.02**
Model (Emo) x ROI (MMPFC)			b=-.48, t=-1.5, p=.14
Model (Emo) x ROI (PC)			b=-.39, t=-1.2, p=.23

Table 1. Statistical Results for Direct Comparisons of Model Fits in Experiments 1 and 2. Full statistics (standardized beta values, t-values, and p-values) for linear mixed-effects regressions comparing the model fit of the planned ToM-relevant models (Condition, Emotion) to the control models (Linguistic, Narrative), and comparing the two ToM-relevant models to each other. Regressions tested for an effect of model (e.g., Condition vs. Linguistic) on the Kendall tau correlation values, which indicate fit to neural RDMs, and included region of interest (ROI) as a covariate. The right temporoparietal junction (RTPJ) was the reference ROI. Regressions also tested for significant Model-by-ROI interactions; non-significant interaction terms were removed from regressions (greyed cells). Significant results at a $p<.05$ threshold are shown in bold text.

3.1.1 Exploratory Model RDMs

Given the initial results, we explored whether a different set of ToM-relevant features would outperform our a priori models. We constructed four new exploratory models: (1) a circumplex model based on valence and arousal features (Barrett, 2006; Russell, 1980), (2) a model based on

abstract appraisal features of events (Barrett, 2006; Russell, 1980) used in a prior study of adults (Skerry & Saxe, 2015), (3) a model that used both emotion and condition features (from the planned models), and (4) a weighted emotion-condition feature model, which was constructed by estimating RDM weights for the three condition labels in addition to the seven emotion features (Khaligh-Razavi et al., 2017). We tested whether these models provided better fits of the neural RDMs than the a priori condition and emotion models, and whether these models were more sensitive to developmental change in ToM responses with age.

3.1.2 Circumplex Model

The circumplex model did not perform above chance in RTPJ (M(SE)=-.01(.01), W=654, p=.92). In the RTPJ and across all ROIs, the circumplex model performed significantly worse than both the condition and emotion models (**Cond vs. Circ:** RTPJ: W=1198, p=.001 (two-tailed); all ROIs: b=-.52, t=-6.0, $p=5.1x10^{-9}$; **Emo vs. Circ:** RTPJ: W=1372, p<.0001; all ROIs: b=-.49, t=-5.9, $p=7.7x10^{-9}$. The circumplex model fit did not differ between children and adults in RTPJ (M(SE) Adult=-.02(.01), Child=.002(.02); effect of age group: b=.23, t=.82, p=.41), or across all ROIs (b=.25, t=.87, p=.39). There was an age group-by-ROI effect such that the circumplex model fit the response in LTPJ (relative to RTPJ) worse in children (b=-.81, t=-2.4, p=.02), and an age-by-ROI (LTPJ)-by-motion interaction (b=.69, t=2.1, p=.04).

3.1.3 Appraisal Model

The appraisal model performed marginally above chance in RTPJ (M(SE)=.02(.01), W=999, p=.09). In the RTPJ, the appraisal model performed worse than the emotion model (W=1136, p=.01), and did not differ significantly from the condition model (W=932, p=.28). Across all ROIs, the appraisal model performed marginally worse than both the condition and emotion models (**Cond vs. App:** b=-.16, t=-1.9, p=.05; **Emo vs. App:** b=-.14, t=-1.7, p=.10). The appraisal model fit did not differ between children and adults in RTPJ (M(SE) Adult=.02(.01), Child=.02(.02); effect of age group: b=-.01, t=-.03, p=.97), or across all ROIs (b=.07, t=.42, p=.68).

3.1.4 Emotion-Condition (EC) Model

Given that the condition and emotion models fit ToM responses best in Experiment 1, we constructed an exploratory model based on both emotion and condition features. This model performed significantly better than chance in RTPJ (M(SE)=.05(.01), W=1316, p=.00005), but did not differ from the a priori ToM-relevant models in RTPJ (**Cond vs. EC:** W=642, p=.20, **Emo vs. EC:** W=695, p=.30) or across all ROIs (**Cond vs. EC:** b=-.0006, t=-.007, p=.99; **Emo vs. EC:** b=.02, t=.23, p=.82, effect of ROI (PC): b=-.23, t=-2.0, p=.04). There was no difference in the fit of the EC model between children and adults in the RTPJ (M(SE) Adult=.05(.01), Child=.04(.02); effect of age group: b=.03, t=.11, p=.91), or across all ROIs (b=-.18, t=-1.1, p=.30).

3.1.5 Weighted Emotion-Condition (WEC) Model

Like the EC model, the WEC model performed significantly better than chance in RTPJ (M(SE)=.04(.01), W=1290, p=.0001), but did not significantly outperform the ToM-relevant models based on condition (W=713, p=.49) or emotion (W=936, p=.39) features alone. The same pattern of results was apparent across all ROIs (**Cond vs. WEC:** b=.12, t=.76, p=.45, model-by-ROI (PC) interaction: b=-.51, t=-2.2, p=.03; **Emo vs. WEC:** b=-.07, t=-.88, p=.38, effect of ROI (PC): b=-.26, t=-2.3, p=.02, no model-by-ROI interactions). The weighted EC model did not outperform the (unweighted) EC model in RTPJ (W=950, p=.33), or across ROIs (b=-.09, t=-1.1, p=.27). There was no difference in the fit of the WEC model between children and adults in the RTPJ (M(SE) Adults=.04(.01), Children=.04(.02); effect of age group: b=.11, t=.40, p=.70), or across all ROIs (b=.09, t=.48, p=.64; ROI (PC)-by-motion interaction: b=.37, t=2.2, p=.03, no other interactions). See Supplementary Figure 5 for a visualization of feature weights, per ROI.

For a visualization of model fits in DMPFC and VMPFC, see Supplementary Figure 6. For a visualization of the model fits per average neural RDM, see Supplementary Figure 8.

Given that the condition and emotion models both outperformed the control models, but did not perform better when combined (despite not being very correlated with one another: r=0.16), in Experiment 2 we used our initial a priori RDMs in confirmatory analyses, and continued to treat the four new RDMs as exploratory.

3.2 Experiment 2

In Experiment 2 we tested the same hypotheses as Experiment 1, in a large sample that included more variability in age and in ToM behavior (n=76 neurotypical and n=29 children diagnosed with an Autism Spectrum Disorder, ages 5-12 years old). Given the results of Experiment 1, our confirmatory hypotheses for Experiment 2 were that 1) emotion and condition models would fit neural responses from ToM brain regions better than chance, and outperform the linguistic and narrative control models, in neurotypical children, 2) the fit of either or both of these models would increase with age or behavioral ToM performance among neurotypical children. Additionally, Experiment 2 enabled us to test if the fit of either or both of these models differed between neurotypical children and children diagnosed with Autism Spectrum Disorder.

3.2.1 Behavioral Battery: Theory of Mind

Children with ASD performed worse on the ToM behavioral task than neurotypical children (M(SE) proportion correct: ASD: .77(.04), NT: .87(.01); effect of group: b=-1.0, t=-5.6, $p=1.7x10^{-7}$), and performance on the task improved with age (b=.40, t=4.3, $p=4.3x10^{-5}$). There was also a group-by-age interaction, such that age had a larger effect on performance in children with ASD, relative to neurotypical children (b=.52, t=2.7, p=.008; see Supplementary Figure 9). Children with ASD also had lower standardized non-verbal IQ scores than neurotypical children (M(SE) ASD: 110(3.3), NT: 117(1.4), effect of group: b=-.48, t=-2.2, p=.03); the same pattern of results for ToM task performance was obtained when additionally controlling for non-verbal IQ.

3.2.2 Model Fits in Neurotypical Children

In the neurotypical child sample (n=76), only the condition model fit the RTPJ neural RDM better than chance (**Cond:** M(SE)=.03(.01), W=2233, p<.00005; **Emo:** .01(.01), W=1633, p=.19; **Ling:** -.003(.01), W=1362, p=.70; **Narr:** -.01(.01), W=949, p=.996). The condition model fit the neural response in the RTPJ of the neurotypical child sample significantly better than both control models (**Cond vs. Ling:** W=2020, p=.002; **Cond vs. Narr:** W=2318, p<.00001), and marginally better than the emotion model (W=1740, p=.08). The emotion model performed better than the narrative control model (W=1946, p=.006), but did not significantly outperform the linguistic control model (W=1715, p=.10). See Figure 3 for visualization of main results and Supplementary Figure 10 for model fits to average neural RDMs.

Across all ROIs, both the condition and emotion models performed significantly better than both control models (see Table 1 for full statistics), and the condition model outperformed the emotion model (Table 1).

Figure 3

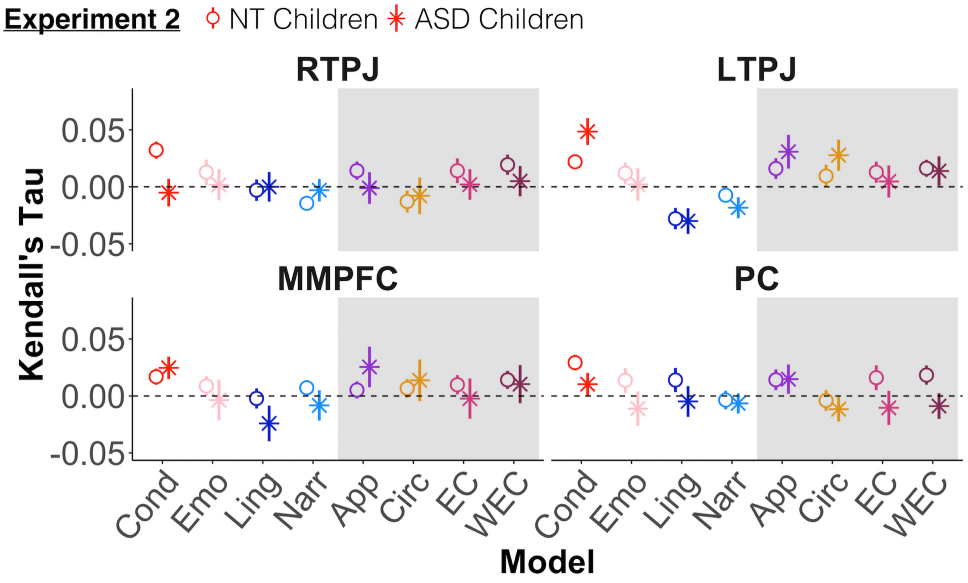


Figure 3. Model Fits in Experiment 2. Plots show the mean Kendall tau correlation (y-axis) between each model (x-axis) and individual neural RDMs, per ROI (RTPJ, LTPJ, PC, MMPFC). Open circles represent means calculated from neurotypical children (n=76); stars represent means calculated from children diagnosed with Autism Spectrum Disorder (ASD; n=29). Lines surrounding mean values indicate standard error from the mean. ToM-relevant (Condition, Emotion) models are shown in red/pink; control (Linguistic, Narrative) models are shown in blue. The shaded area indicates exploratory models, which included a model based on abstract appraisal features (App), a circumplex model based on valence and arousal (Circ), and models that included both emotion and condition features (EC, W (weighted) EC)).

Does the extent to which ToM-relevant models fit neural activity in ToM brain regions vary with age or ToM behavioral score? In neurotypical children, the fit of the condition model increased with age in the RTPJ (effect of age: b=.29, t=2.6, p=.01, effect of motion: b=-.06, t=-.49, p=.6), and across all ROIs (effect of age: b=.31, t=2.8, p=.007, effect of motion: b=-.03, t=-.41, p=.68, effects of ROIs: bs<|.26|, ts<|1.9|, ps>.06, age-by-ROI (PC) interaction: b=-.28, t=-2.1, p=.04, other age-by-ROI interactions: bs<|.27|, ts<|2.0|, ps>.05, no other interactions; Figure 5). The effect of age on the condition model fit remained significant with a Bonferroni correction for multiple comparisons (two tests; α=.025).

The emotion model fit did not change with age in the RTPJ (effect of age: b=.01, t=.13, p=.90, effect of motion: b=-.28, t=-2.4, p=.02) or across all ROIs (effect of age: b=.02, t=.14, p=.89, effects of ROIs: bs<|.06|, ts<|.4|, ps>.7, effect of motion: b=-.30, t=-2.6, p=.01, age-by-ROI (PC)-by-motion interaction: b=-.26, t=-2.1, p=.04; all other interactions were not significant).

There was no significant correlation between the fit of the condition or emotion model fits and ToM behavioral score, either in RTPJ (**Cond:** b=-.07, t=-.50, p=.62; **Emo:** b=.18, t=1.4, p=.17, controlling for age and motion), or across all ROIs (**Cond:** b=-.02, t=-.25, p=.81; **Emo:** effect of ToM: b=.003, t=.04, p=.97).

Because overall accuracy on the fMRI behavioral task (which was orthogonal to ToM processes, and served to ensure attention to the stories) increased with age among neurotypical children in Experiment 2, in post-hoc analyses we confirmed that the effect of age on the condition model fit remained significant when additionally controlling for accuracy (**RTPJ:** effect of age: b=.29, t=2.2, p=.03, effect of accuracy: b=-.02, t=-.12, p=.91; **all ROIs:** effect of age: b=.29, t=2.3, p=.02, effect of accuracy: b=.04, t=.43, p=.67).

3.2.3 Exploratory Searchlight Analyses in Neurotypical Children

To ensure that we did not miss unpredicted effects in other brain regions, we conducted a whole-brain searchlight analysis across all neurotypical children from Experiments 1 and 2 (n=96), in order to discover brain regions in which response patterns correlated with the condition and emotion models. The searchlight analysis revealed that the condition model uniquely predicted response patterns in ToM brain regions (see Figure 4 for visualization, and Supplementary Table 3 for details of results). While there were not any significant clusters predicted by the emotion model, small clusters in the right superior temporal sulcus and premotor cortex were present at more lenient statistical thresholds (p<.001, k=10, uncorrected; see Supplementary Figure 11).

Figure 4

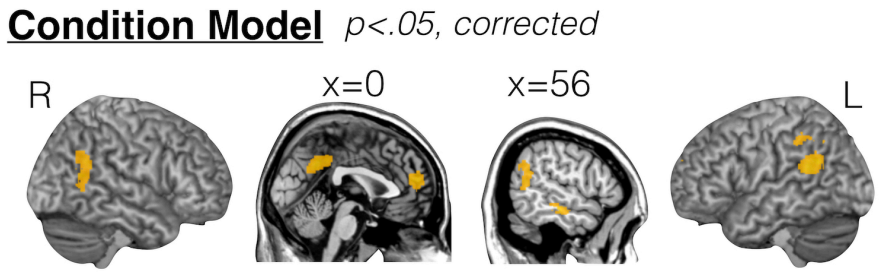


Figure 4. Searchlight Analysis for Condition Model Fit. An exploratory searchlight analysis revealed that response patterns in ToM brain regions correlated with the condition model. Results have been corrected for multiple comparisons (p<.05, SnPM). See Supplementary Table 3 for detailed information about the significant clusters.

3.2.4 Model Fits in Children Diagnosed with Autistic Spectrum Disorder

In the ASD child sample (n=29), none of the model RDMs fit the RTPJ RDM better than chance (**Cond:** M(SE)=-.005(.01), W=180, p=.80; **Emo:** .002(.01), W=228, p=.42; **Ling:** -.00003(.01), W=206, p=.60; **Narr:** -.003(.01), W=190, p=.72). The condition and emotion models did not outperform the two control models in RTPJ (**Cond vs. Ling:** W=212, p=.55; **Cond vs. Narr:** W=221, p=.47; **Emo vs. Ling:** W=234, p=.37, **Emo vs. Narr:** W=249, p=.25); there was additionally no difference in the fit of the condition and emotion models to the RTPJ RDM (W=192, p=.71); Figure 3.

Across all ROIs, the condition and emotion models did not outperform the control models, and the condition model did not outperform the emotion model. Interestingly, there were significant model-by-ROI interactions such that the condition model fit the neural data significantly better than the linguistic and narrative control models in the LTPJ and MMPFC, relative to the RTPJ (see Table 1 for full statistics and Figure 5 for visualization).

3.2.5 Direct Comparisons Between Neurotypical Children and Children with an ASD Diagnosis

We directly compared the fit of the condition and emotion models to the neural data across children with and without a diagnosis of ASD. The condition model fit the RTPJ responses in neurotypical children significantly better than children diagnosed with ASD (M(SE) NT=.03(.01), ASD=-.005(.01); effect of group: b=-.56, t=-2.7, p=.009, effect of motion: b=-.17, t=-1.8, p=.08, no group-by-motion interaction; Figure 5). The significant group difference for the condition model fit remained significant with a Bonferroni correction for multiple comparisons

(two tests; α=.025), and when additionally controlling for non-verbal IQ (effect of group: b=-.56, t=-2.5, p=.015, effect of IQ: b=.06, t=.54, p=.59, effect of motion: b=-.14, t=-1.5, p=.15). The fit of the emotion model to RTPJ responses did not differ across children with and without ASD (M(SE) NT=.01(.01), ASD=.002(.01); effect of group: b=-.13, t=-.61, p=.54, effect of motion: b=-.21, t=-2.1, p=.03, no group-by-motion interaction).

Across all ToM ROIs, there was a significant effect of group such that the condition model fit the neural RDMs better in neurotypical children, relative to children diagnosed with ASD (effect of group (ASD): b=-.62, t=-2.9, p=.005; effect of ROIs: bs<|.26|, ts<|1.9|, ps>.06, effect of motion: b=-.12, t=-1.8, p=.08). Additionally, significant group-by-ROI interactions indicated that specifically in the ASD group, the fit of the condition model to the RTPJ was worse than the fit of this model to the LTPJ and MMPFC (group-by-ROI (LTPJ) interaction: b=1.0, t=4.1, p=.00005, group-by-ROI (MMPFC) interaction: b=.74, t=2.9, p=.004; group-by-ROI (PC) interaction: b=.30, t=1.2, p=.24). The significant group difference for the condition model fit remained significant with a Bonferroni correction for multiple comparisons (two tests for two ToM models; α=.025), and the same pattern of results was obtained when additionally controlling for non-verbal IQ. There was no effect of ASD diagnosis on the model fit of the emotion model across ROIs (effect of group: b=-.13, t=-.62, p=.54, effect of motion: b=-.31, t=-2.7, p=.007, effects of ROIs: bs<|.05|, ts<|.35|, ps>.7, group-by-ROI interactions: bs<|.17|, ts<|.7|, ps>.5, ROI (PC)-by-motion interaction: b=.34, t=2.5, p=.01, ROI (LTPJ)-by-group-by-motion interaction: b=-.67, t=-2.6, p=.01, ROI (PC)-by-group-by-motion interaction: b=-.57, t=-2.2, p=.03, ROI (MMPFC)-by-group-by-motion interaction: b=-.85, t=-3.3, p=.001, all other interactions were not significant).

Given that we observed reduced fit of the condition model to the RTPJ of children with ASD, we conducted exploratory analyses to test whether any of the other models showed a better fit in this group. We did not find any evidence for a model that fit the neural data better in ASD (all group effects: bs<|.21|, ts<|1|, ps>.3).

3.2.6 Examination of the Contribution of Univariate Responses

Our results suggest that multivariate analyses were sensitive to developmental change with age in the neurotypical children, and to differences between neurotypical children and children diagnosed with ASD. To address the possibility that these results primarily reflect differences in univariate responses, we conducted supplementary analyses to test 1) whether similar results were obtained with multivariate analyses that used a dissimilarity metric that is insensitive to the univariate response (Pearson correlation distance, rather than Euclidean distance; Walther et al., 2016), and 2) whether a univariate measure – response selectivity – was also sensitive to individual differences in ToM responses. Response selectivity was calculated as the magnitude of response (average beta) of the mental condition minus the magnitude of response to the social condition, in the same ROIs used for multivariate analyses (following the pre-registered procedure for group ROIs; https://osf.io/wzd8a).

Overall, we observed similar results from multivariate analyses that used Pearson correlation distance (see Supplementary Materials and Supplementary Table 2 for full statistics, and Supplementary Figure 7 for visualization of model fits). Additionally, the univariate measure of response selectivity did not increase significantly with age among neurotypical children, and did not differ between neurotypical children and children with ASD (Supplementary Figure 12). Together, these results suggest that the multivariate approach captures individual differences in ToM responses that may not be detectable with univariate approaches.

Figure 5

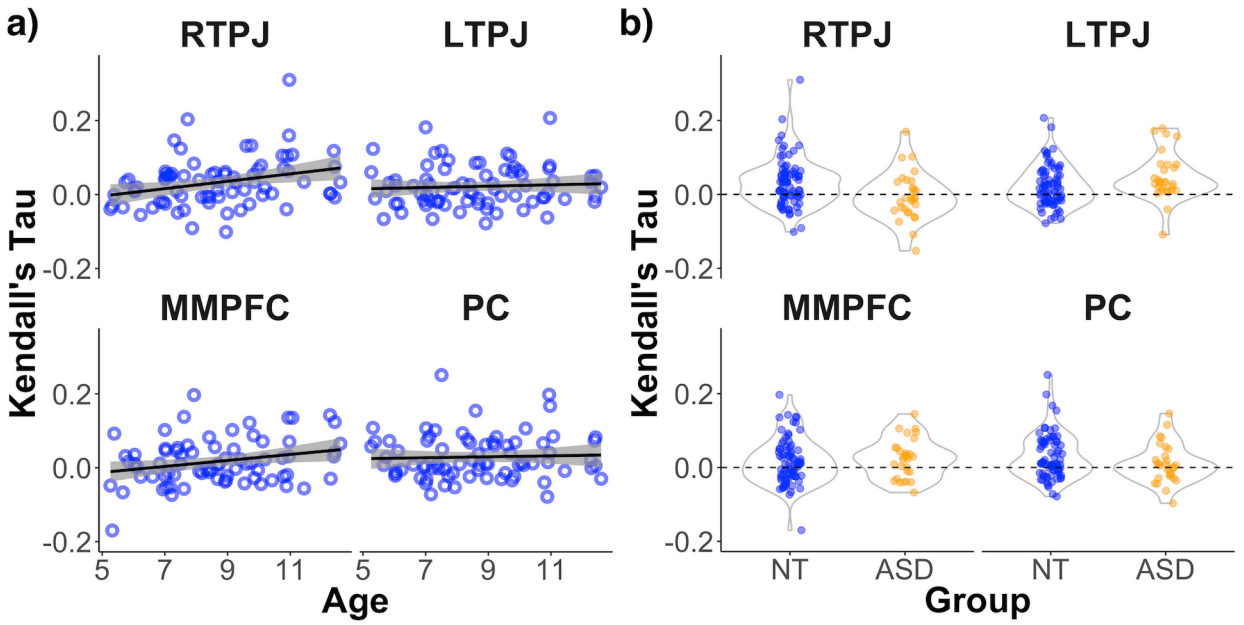


Figure 5. Condition Model Fit per ROI (Experiment 2). All plots show Kendall's tau correlation values (y-axis) calculated between individual neural RDMs (per ROI) and the condition model by **a)** age (in years, x-axis), among neurotypical children (n=76, blue) and by **b)** group (x-axis; children diagnosed with Autism Spectrum Disorder (ASD, n=29) are shown in orange).

4. Discussion

Almost all fMRI studies of theory of mind, and all fMRI studies of theory of mind in children, have used univariate analyses in order to characterize response magnitude and selectivity in ToM brain regions. Multivariate pattern analyses have the potential to describe the development of the structure of representational content within a stimulus category (e.g., mental states). This kind of description of neural responses may be particularly important for capturing developmental change or differences in the structural organization of ToM concepts: i.e., individual differences in sensitivity to the conceptual distinctions and causal relationships between mental states. Here, we show that condition labels and emotion features capture some of the pattern of activity in ToM brain regions in children. Moreover, neural responses become increasingly organized by condition labels with age in a relatively large sample of neurotypical children. Additionally, condition labels do not appear to organize neural response patterns in the RTPJ in children diagnosed with Autism Spectrum Disorder. These results suggest that there are real, stable features that organize neural responses in ToM brain regions in children, and that multivariate analyses can be used to measure developmental change and differences in the conceptual structure of theory of mind representations in childhood.

Developmental change in ToM representations in childhood was best captured by the condition label model, which indicated whether a story involved descriptions of mental states (e.g., beliefs, desires, emotions), general social information (e.g., personality traits, appearance, or enduring relationships), or just descriptions of causal events in the world (e.g., a tree growing fruit; a bird laying eggs). This model marked the extent to which a story was about the mind, and captured a distinction between preferred (Mental) and non-preferred stimuli (Social, Physical) for ToM brain regions typically characterized by univariate measures. Prior fMRI studies of adults have provided similar evidence for distinct response patterns for preferred and non-preferred stimuli. For example, ToM brain regions have distinct neural response patterns for stories that describe characters' mental states versus stories that describe physical events (Koster-Hale et al., 2017). Here, response patterns in ToM brain regions were increasingly organized according to the condition label model among children, and remained organized by this model in adults.

The condition model was not only sensitive to developmental change in ToM neural response patterns, but it was also sensitive to differences between neurotypical children and children diagnosed with Autism Spectrum Disorder. Specifically, the condition model did not predict RTPJ response patterns in children with ASD, and the condition model fit in RTPJ was significantly worse in children with ASD, compared to neurotypical children. Given our small sample of children with ASD (n=29) and the heterogeneity of social deficits in this disorder (Byrge et al., 2015; Lombardo et al., 2016; Pierce et al., 2016), we report this result with caution. Nonetheless, there are a few notable aspects of our approach that strengthen our findings. First, we measured functional responses in children with ASD during social processing – which is inherently challenging and not frequently done. Second, participant motion and data quality were matched between neurotypical children and children with ASD. Third, we had a relatively large sample of neurotypical children to use as a comparison group. And finally, we observed some specificity in the group difference: that is, the model fits in LTPJ and MMPFC in the ASD sample look similar to those in the NT sample. If the observed group difference was driven by (non-significant) differences in data quality or an undetected confound, it is difficult to explain why response patterns in RTPJ alone were disrupted in children with ASD. On the other hand, it is possible that response patterns in RTPJ would be particularly sensitive to the social cognitive

differences between NT and ASD children, given its role in ToM reasoning (Gweon et al., 2012; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010a), and prior evidence for disrupted patterns in RTPJ in adults with ASD (Koster-Hale et al., 2013). This result raises several questions for future research. In particular, given that none of the planned or exploratory models provided a good description of RTPJ response patterns in children with ASD, what dimensions do predict these response patterns?

A key goal of the current study was to go beyond the condition labels, and describe the structure of representations *within* the category of mental states (beliefs, desires, emotions). In addition to the condition label model, we constructed an emotion feature model based on ratings to seven emotions. Like the condition model, the emotion model fit neural responses in ToM brain regions better than models based on linguistic and narrative features. In exploratory analyses, we compared these two models to an abstract event appraisal model derived directly from a prior fMRI study of emotion representations in adults (Skerry & Saxe, 2015). In that study, the abstract event appraisal model outperformed a model based on the six basic emotions (Cohen et al., 2017; Du et al., 2014; Ekman, 1992) and a circumplex model based on valence and arousal (Barrett, 2006; Russell, 1980) in predicting behavioral and response similarity in ToM brain regions to verbal narratives (Skerry & Saxe, 2015). Here, the condition and emotion models both fit neural responses in ToM brain regions better than the abstract event appraisal model. One intriguing hypothesis is that responses in ToM brain regions transition from being organized by condition and emotion features in childhood, to abstract event appraisals in adults. Our evidence does not provide support for this hypothesis – the appraisal model fit did not increase between childhood and adulthood (Experiment 1), and the fit of the condition model actually increased with age among children (Experiment 2). An alternative explanation is that this difference in results is due to the methodological constraints of the current experiment. The appraisal model RDM characterizing the story stimuli in the current experiment was most correlated with the circumplex model RDM – which characterized the stories along valence and arousal dimensions only. Experimental stimuli necessarily constrain the extent to which different features can explain variance in neural responses: if the stories did not vary in the extent to which they evoked abstract event appraisals, then these features cannot predict differences in neural responses across stories. Thus, while the exploratory test of whether abstract event appraisal

features organize responses in ToM brain regions in childhood could be considered a strong test of the generalization of these features to other stimuli and populations, this test was likely underpowered given our experimental design. Additional research is needed to further explore developmental change in the structure of mental state representations.

Our results suggest that multivariate approaches are promising for characterizing individual differences in the structure and content of mental state representations in children, but they also suggest the need for future studies that are designed with these specific analyses in mind. The relatively low model fits suggest that these analyses were up against the limitations imposed by the experimental design — i.e., the use of complex stimuli, the use of just a single presentation of each item for pattern analyses, and more generally the use of very little data per participant. These limitations are in stark contrast to most multivariate fMRI studies in adults, which typically measure response patterns across several repetitions of the same stimulus, or at least several stimuli per stimulus feature or category. Because the multivariate results overall look very similar in adults and children (Experiment 1), the small correlations observed between neural and model RDMs are likely due to experimental limitations, rather than reflecting challenges specific to the pediatric data. While collecting a large amount of data within individual child participants is inherently difficult, future studies may benefit from developing stimuli that target specific features or dimensions of mental states (Koster-Hale et al., 2013; Koster-Hale et al., 2014; 2017).

A key benefit of multivariate analyses is that they characterize features that drive neural response (dis)similarity. In the domain of theory of mind, this benefit carries particular weight. Theory of mind development involves refining distinctions between ToM concepts, and constructing an increasingly sophisticated, flexible theory about the causal relationships between them. Multivariate approaches provide a way to capture these structural changes in ToM representations – which may be key for characterizing development and disorders in ToM.

Acknowledgements
We gratefully acknowledge support by the Ellison Medical Foundation, and by a NSF Graduate Research Fellowship (#1122374 to H.R.). We thank the Athinoula A. Martinos Imaging Center

at the McGovern Institute for Brain Research at MIT, Dorit Kliemann and Todd Thompson for helpful advice on analyses, Marina Bedny, Swetha Dravida and Mika Asaba for help with data collection and organization, and the individuals who participated.

References

Adolphs, R. (2009). The Social Brain: Neural Basis of Social Knowledge. *Annual Review of Psychology*, *60*(1), 693–716. http://doi.org/10.1146/annurev.psych.60.110707.163514

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119.

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (DSM-5(R)). American Psychiatric Pub.

Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. *Understanding Other Minds: Perspectives From Developmental Cognitive Neuroscience*, *2*, 3–20.

Barrett, L. F. (2006). Valence is a basic building block of emotional life. *Journal of Research in Personality*, *40*(1), 35–55.

Bedny, M., Richardson, H., & Saxe, R. (2015). "Visual" Cortex Responds to Spoken Language in Blind Children. *Journal of Neuroscience*, *35*(33), 11674–11681. http://doi.org/10.1523/JNEUROSCI.0634-15.2015

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101.

Bradmetz, J., & Schneider, R. (1999). Is Little Red Riding Hood afraid of her grandmother? Cognitive vs. emotional response to a false belief. *The British Journal of Developmental Psychology*, *17*(4), 501–514.

Bruneau, E. G., Pluta, A., & Saxe, R. (2012). Distinct roles of the "shared pain'and 'theory of mind' networks in processing others" emotional suffering. *Neuropsychologia*, *50*(2), 219–231.

Byrge, L., Dubois, J., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. *Journal of Neuroscience*, *35*(14), 5837–5850.

Carp, J. (2013). Optimizing the order of operations for movement scrubbing: Comment on Power et al. *NeuroImage*, *76*, 436–438.

Carrington, S. J., & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping*, *30*(8), 2313–2335. http://doi.org/10.1002/hbm.20671

Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A Distinct Role of the Temporal-Parietal Junction in Predicting Socially Guided Decisions. *Science*, *337*(6090), 109–111. http://doi.org/10.1126/science.1219681

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., et al. (2017). Computational approaches to fMRI analysis. *Nature Publishing Group*, *20*(3), 304.

Coutanche, M. N., Thompson-Schill, S. L., & Schultz, R. T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage*, *57*(1), 113–123. http://doi.org/10.1016/j.neuroimage.2011.04.016

de Bie, H. M. A., Boersma, M., Wattjes, M. P., Adriaanse, S., Vermeulen, R. J., Oostrom, K. J., et al. (2010). Preparing children with a mock scanner training protocol results in high quality structural and functional MRI scans. *European Journal of Pediatrics*, *169*(9), 1079–1085. http://doi.org/10.1007/s00431-010-1181-z

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, *43*(1), 44–58.

Döhnel, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, M. (2012). Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage*, *60*(3), 1652–1661.

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), E1454–E1462.

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., et al. (2013). Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE*, *8*(9), e75468. http://doi.org/10.1371/journal.pone.0075468

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*(3), 550-553.

Ellsworth, P. C. (2013). Appraisal theory: Old and new questions. *Emotion Review*, *5*(2), 125–131.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678.

Gilbert, S. J., Meuwese, J. D., Towgood, K. J., Frith, C. D., & Burgess, P. W. (2009). Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: a multi-voxel similarity analysis. *Brain*, *132*(4), 869–878.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1-2), 145–171.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, *100*(1), 253–258.

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of Mind Performance in Children Correlates With Functional Specialization of a Brain Region for Thinking About Thoughts. *Child Development*, *83*(6), 1853–1868. http://doi.org/10.1111/j.1467-8624.2012.01829.x

Hallquist, M. N., Hwang, K., & Luna, B. (2013). The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. *NeuroImage*, *82*, 208–225.

Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition & Emotion*, *3*(4), 379–400.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2013). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex (New York, N.Y. : 1991)*, *24*(8), 1979–1987.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Jastorff, J., Huang, Y. A., Giese, M. A., & Vandenbulcke, M. (2015). Common neural correlates of emotion perception in humans. *Human Brain Mapping*, *36*(10), 4184–4201.

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, *83*, 201–226.

Kaufman, A. S. (1997). KBIT-2: Kaufman Brief Intelligence Test. Minneapolis, MN: NCS Pearson.

Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., et al. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain.

Nature Publishing Group, *22*(10), 1617–1623.
Keil, B., Alagappan, V., Mareyam, A., McNab, J. A., Fujimoto, K., Tountcheva, V., et al. (2011). Size-optimized 32-channel brain arrays for 3 T pediatric imaging. *Magnetic Resonance in Medicine*, *66*(6), 1777–1787.
Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2017). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, *76*, 184–197.
Kim, J., Schultz, J., Rohe, T., Wallraven, C., Lee, S.-W., & Bülthoff, H. H. (2015). Abstract representations of associated emotions in the human brain. *Journal of Neuroscience*, *35*(14), 5655–5663.
Kliemann, D., Richardson, H., Anzellotti, S., Ayyash, D., Haskins, A. J., Gabrieli, J. D., & Saxe, R. R. (2018). Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without Autism. *Cortex*, *103*, 24–43.
Koster-Hale, J., Bedny, M., & Saxe, R. (2014). Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition*, *133*(1), 65–78. http://doi.org/10.1016/j.cognition.2014.04.006
Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, *161*, 9–18. http://doi.org/10.1016/j.neuroimage.2017.08.026
Koster-Hale, J., Saxe-R. (2013) Theory of Mind: A Neural Prediction Problem. (2013). *Neuron, 79*(5), 836–848. http://doi.org/10.1016/j.neuron.2013.08.020
Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, *110*(14), 5648–5653.
Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. http://doi.org/10.1016/j.tics.2013.06.007
Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.
Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., et al. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience*, *22*(7), 1623–1635.
Lombardo, M. V., Lai, M.-C., Auyeung, B., Holt, R. J., Allison, C., Smith, P., et al. (2016). Unsupervised data-driven stratification of mentalizing heterogeneity in autism. *Scientific Reports*, *6*, 35333.
Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., et al. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223.
McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). Coh-Metrix version 3.0. *Retrieved [4/1/15] From Http://Cohmetrix. Com*.
Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118.
Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the

medial prefrontal cortex to knowledge about mental states. *NeuroImage*, *28*(4), 757–762.
Mitchell, T.M., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine learning, 57*(1-2), 145-175.
Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Sert, du, N. P., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.
Nelson, N. L., Widen, S. C., & Russell, J. A. (2006). Children's understanding of emotions' causes and consequences: Labeling facial expression and stories. Presented at the Poster presented at the 18th Annual American Psychological Society Convention, New York, NY.
Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005
Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal Representations of Perceived Emotions in the Human Brain. *Journal of Neuroscience*, *30*(30), 10127–10134. http://doi.org/10.1523/JNEUROSCI.2161-10.2010
Pelphrey, K. A., Shultz, S., Hudac, C. M., & Vander Wyk, B. C. (2011). Research review: constraining heterogeneity: the social brain and its development in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *52*(6), 631–644.
Pelphrey, K., Adolphs, R., & Morris, J. P. (2004). Neuroanatomical substrates of social cognition dysfunction in autism. *Mental Retardation and Developmental Disabilities Research Reviews*, *10*(4), 259–271.
Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1), S199–S209. http://doi.org/10.1016/j.neuroimage.2008.11.007
Pierce, K., Marinero, S., Hazin, R., McKenna, B., Barnes, C. C., & Malige, A. (2016). Eye tracking reveals abnormal visual preference for geometric images as an early biomarker of an autism spectrum disorder subtype associated with increased symptom severity. *Biological Psychiatry*, *79*(8), 657–666.
Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, *1*(2), 127–152.
Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental Psychology*, *33*(1), 12.
Richardson, H. (2019). Development of Brain Networks for Social Functions: Confirmatory Analyses in a Large Open Source Dataset. *Developmental Cognitive Neuroscience, 37,* 100598.
Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), 1027.
Ruffman, T., & Keenan, T. R. (1996). The belief-based emotion of suprise: The case for a lag in understanding relative to false belief. *Developmental Psychology*, *32*(1), 40.
Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.
Sabbagh, M. A., Bowman, L. C., Evraire, L. E., & Ito, J. M. B. (2009). Neurodevelopmental correlates of theory of mind in preschool children. *Child Development*, *80*(4), 1147–1162. http://doi.org/10.1111/j.1467-8624.2009.01322.x
Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for

perceiving and reasoning about other people in school-aged children. *Child Development*, *80*(4), 1197–1209.
Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842.
Saxe, R., & Powell, L. J. (2006). It's the thought that counts specific brain regions for one component of theory of mind. *Psychological Science*, *17*(8), 692–699.
Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–1399. http://doi.org/10.1016/j.neuropsychologia.2005.02.013
Scherer, K. R. (1999). Appraisal theory. *Handbook of Cognition and Emotion*, 637–663.
Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *34*(48), 15997–16008. http://doi.org/10.1523/JNEUROSCI.1676-14.2014
Skerry, A. E., & Saxe, R. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology*, *25*(15), 1945–1954. http://doi.org/10.1016/j.cub.2015.06.009
Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*(2), 204–216.
Spunt, R. P., Kemmerer, D., & Adolphs, R. (2015). The neural basis of conceptualizing the same action at different levels of abstraction. *Social Cognitive and Affective Neuroscience*, nsv084.
Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, *113*(1), 194–199. http://doi.org/10.1073/pnas.1511905112
Thesen, S., Heid, O., Mueller, E., & Schad, L. R. (2000). Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magnetic Resonance in Medicine*, *44*(3), 457–465.
Thornton, M. A., & Mitchell, J. P. (2017a). Consistent neural activity patterns represent personally familiar people. *Journal of Cognitive Neuroscience*, *29*(9), 1583–1594.
Thornton, M. A., & Mitchell, J. P. (2017b). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, *28*(10), 3505–3520.
Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The social brain automatically predicts others' future mental states. *Journal of Neuroscience*, *39*(1), 140–148.
Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200.
Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541.
Whitfield-Gabrieli, S., Nieto-Castanon, A., & Ghosh, S. (2011). Artifact Detection Tools (ART). *Cambridge, MA. Release Version*, *7*(19), 11.
Widen, S. C. (2016). The development of children's concepts of emotion. *Handbook of Emotions, Eds Barrett LF, Lewis M, Haviland-Jones JM (Guilford, New York)*, 307–318.
Wu, Y., & Schulz, L. E. (2018). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child Development*, *89*(2), 649–662.
Xiao, Y., Geng, F., Riggins, T., Chen, G., & Redcay, E. (2019). Neural correlates of developing theory of mind competence in early childhood. *NeuroImage*, *184*, 707–716.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010a). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*(15), 6753–6758. http://doi.org/10.1073/pnas.0914826107

Young, L., Nichols, S., & Saxe, R. (2010b). Investigating the Neural and Cognitive Basis of Moral Luck: It's Not What You Do but What You Know. *Review of Philosophy and Psychology*, *1*(3), 333–349. http://doi.org/10.1007/s13164-010-0027-y

Supplementary Materials

Table of Contents

Supplementary Tables

Supplementary Figures

Departures from Pre-Registered Analyses

1. Regions of Interest

Based on prior studies relating neural development to theory of mind in children (Sabbagh et al., 2009; Gweon et al., 2012), we planned to conduct primary analyses in dorso-medial prefrontal cortex (DMPFC) and right temporoparietal junction (RTPJ), and to conduct exploratory analyses in other ToM brain regions. Upon calculating the noise ceiling for each ToM brain region and sample, we found that we could not reliably estimate model fits to data extracted from DMPFC and VMPFC (see Supplementary Figure 3). Thus, subsequent statistical analyses were not conducted for these two regions. See Supplementary Figures 6 and 8 for a visualization of model fits in these regions.

2. Motion

We pre-registered testing whether number of artifact timepoints correlated with mean translation (calculated prior to artifact timepoint removal), and using number of artifact timepoints as a covariate in regressions testing for between-subject effects. Number of artifact timepoints was correlated with mean translation in both experiments (rs>.57). We opted to use mean translation rather than number of artifact timepoints as our motion covariate in regressions for two reasons: (1) as pointed out by an anonymous reviewer, this measure better captures individual differences in motion, as it is sensitive to small movements (that don't reach the threshold of being artifact timepoints), and (2) the number of artifact timepoints metric is affected by the number of runs included, which varies across participants.

3. Features for Weighted Feature Model

We planned to calculate feature weights for the 7 emotion model features if the emotion model outperformed the condition model as well as the control models. Because the emotion model did not outperform the condition model, we calculated feature weights for the 3 condition and 7 emotion features, and constructed a weighted emotion condition (WEC) model (Supplementary Figure 5).

We also initially planned to test for significant change with age in dimension weights, per feature. We opted against conducting this analysis because it would require estimating feature weights using a single subject's neural RDM. Given the limited amount of data per participant, we instead estimated feature weights using the average neural RDMs from the full Experiment 1 participant sample (n=57 children and adults), per ROI (Supplementary Figure 5).

4. Addition of Comparison Between Neurotypical Children and Children with ASD

Finally, it is worth noting that the pre-registered analyses were designed with the large neurotypical sample in mind. Tests for differences between neurotypical children and children diagnosed with autism were not pre-registered.

4. Tests for Developmental Change

We initially planned to test for developmental change in the overall fit of the emotion model to the neural data if the emotion model outperformed the condition model as well as the control models. Because the emotion model did not outperform the condition model, we tested for developmental change in both of the ToM-relevant models (condition and emotion).

Finally, we planned to test whether the model fit of ToM-relevant models change with age more than control models using William's r tests ("psych" package in R). In RTPJ, the increase in model fit with age did not differ from change with age in other model fits (Emotion: z=1.5, p=.13; Linguistic: z=1.9, p=.06, Narrative: z=1.5, p=.13). We additionally used mixed effects linear regressions to run the same test across all regions of interest. Across all regions, the condition model fit increased more with age than the linguistic (b=-.17, t=-2.3, p=.02) and narrative (b=-.16, t=-2.1, p=.04) model fits; this model x age interaction was not significant in the comparison with the emotion model fit (b=-.12, t=-1.5, p=.12).

Planned Analysis: Within-Condition Response Pattern Similarity

In our analysis plan, we hypothesized that the pairwise response dissimilarity between Mental stories would increase with age - i.e., that responses to distinct stories describing mental states would become less similar to one another across childhood. We did not find significant change with age in the dissimilarity of neural responses between Mental stimuli in either experiment (Experiment 1: effect of age group: b=.10, t=.35, p=.73; Experiment 2: effect of age (continous): b=.20, t=1.7, p=.10); see Supplementary Figure 13.

Results Using Pearson Correlation Dissimilarity Metric

In our pre-registered analyses, we calculated neural response dissimilarity between items using Euclidean distance. We were primarily interested in the relative fits of our model RDMs to the neural RDMs, and didn't have specific hypotheses about the relative role of univariate and multivariate signals in neural response dissimilarity. Here, we present results of the primary analyses from each experiment using Pearson correlation distance as the neural response dissimilarity metric, which (unlike Euclidean distance) is insensitive to variation in the univariate response. Overall, we observe a similar pattern of results in the relative fits of the neural RDMs to the models (see Supplementary Figure 7).

Experiment 1

In the RTPJ, patterns of neural activity were correlated with both the condition and emotion models, significantly better than chance (Chance = 0, **Cond:** M(SE)=.02(.01), W=1068, p=.03; **Emo:** .02(.01), W=1035, p=.049; Wilcoxon Signed-rank tests); the linguistic and narrative control models did not differ significantly from chance (**Ling: -**.004(.01), W=702, p=.84; **Narr:** -.01(.01), W=528, p=.99). The emotion and condition models did not differ in their fit to the RTPJ RDM (W=818, p=.95; two-tailed paired). The condition and emotion models each performed significantly better than either control model (**Cond vs. Ling:**

W=1039, p=.046; **Cond vs. Narr:** W=1225, p=.0008; **Emo vs. Ling:** W=1079, p=.02; **Emo vs. Narr:** W=1195, p=.002).

Across all ToM ROIs, the condition model outperformed both control models as well as the emotion model. The emotion model performed significantly better than the narrative control model (mixed effects linear regressions, see Supplementary Table 2 for full statistics).

We then tested for group differences in the fit of the two ToM-relevant feature models, based on age group (adult vs. child). In the RTPJ, there were no differences in the fit of either model (**Cond:** M(SE) Adult=.02(.01), Child=.02(.02); **Emo:** Adult=.03(.01), Child =-.002(.02); effects of age group: bs<.4, ts<1.4, ps>.15, controlling for motion; no group-by-motion interactions). Similarly, there was no effect of age group on the fit of these models across all ROIs (**Cond:** effect of age group: b=.09, t=.32, p=.75, effects of ROIs: bs<|.20|, ts<|.9|, ps>.3, effect of motion: b=-.04, t=-.22, p=.83, age-by-ROI (MMPFC)-by-motion interaction: b=.78, t=2.1, p=.03, no other significant interactions; **Emo:** effect of age group: b=-.36, t=-1.3, p=.19, effect of ROI (LTPJ): b=-.03, t=-.13, p=.90, effect of ROI (MMPFC): b=-.61, t=-3.0, p=.003, effect of ROI (PC): b=-.41, t=-2.0, p=.04, effect of motion: b=-.11, t=-.59, p=.56, age-by-ROI (MMPFC) interaction: b=1.0, t=3.0, p=.003), age-by-ROI (PC) interaction: b=.96, t=2.8, p=.006), ROI (PC)-by-motion interaction: b=-.50, t=-2.2, p=.03, age-by-ROI-by-motion interaction: b=.65, t=2.0, p=.049; no other significant interactions). See Supplementary Figure 7 for visualization.

Experiment 2: Neurotypical Children

In the neurotypical child sample (n=76), only the condition model fit the RTPJ neural RDM better than chance (**Cond:** M(SE)=.03(.01), W=2090, p<.0006; **Emo:** -.002(.01), W=1388, p=.65; **Ling:** -.005(.01), W=1203, p=.91; **Narr:** -.009(.01), W=1212, p=.9). The condition model fit the neural response in the RTPJ of the neurotypical child sample significantly better than both control models (**Cond vs. Ling:** W=2086, p=.0006; **Cond vs. Narr:** W=2238, p=.00003), and better than the emotion model (W=2119, p=.0003). The emotion model did not significantly outperform the linguistic (W=1412, p=.6) or narrative (W=1606, W=.23) control models (Supplementary Figure 7).

Across all ROIs, the condition model alone performed significantly better than both control models, and also significantly outperformed the emotion model (Supplementary Table 2).

Change with Age and Behavioral Theory of Mind

In neurotypical children, the fit of the condition and emotion models did not change with age in RTPJ (**Cond:** effect of age: b=.09, t=.77, p=.45, effect of motion: b=-.19, t=-1.7, p=.10; **Emo:** effect of age: b=.02, t=.17, p=.87, effect of motion: b=-.06, t=-.54, p=.59), or across all ROIs (**Cond:** effect of age: b=.08, t=1.1, p=.27, effects of ROIs: bs<|.15|, ts<|.97|, ps>.3, effect of motion: b=-.20, t=-1.8, p=.08, ROI (MMPFC)-by-motion interaction: b=.35, t=2.4, p=.02, ROI (LTPJ)-by-motion interaction: b=.32, t=2.1, p=.03, no other significant interactions; **Emo:** effect of age: b=-.05, t=-.67, p=.50, effects of ROIs: bs<|.12|, ts<|.75|,

ps>.4, effect of motion: b=-.08, t=-.72, p=.48, ROI (MMPFC)-by-motion interaction: b=.40, t=2.6, p=.009, no other significant interactions).

ToM behavioral score was not correlated with condition or emotion model fit in RTPJ (**Cond:** b=.15, t=1.1, p=.26; **Emo:** b=.15, t=1.1, p=.28, controlling for age and motion), or across all ROIs (**Cond:** b=.11, t=1.3, p=.19; **Emo:** effect of ToM: b=.18, t=1.5, p=.15, controlling for age and motion; no significant ToM-by-ROI interactions).

Experiment 2: Children Diagnosed with Autistic Spectrum Disorder
As in the primary results (using Euclidean Distance), none of the model RDMs fit the RTPJ RDM better than chance in the ASD child sample (n=29; **Cond:** M(SE)=.001(.01), W=223, p=.46; **Emo:** .009(.01), W=252, p=.23; **Ling:** .008(.01), W=204, p=.62; **Narr:-**.01(.01), W=145, p=.94). The condition and emotion models did not outperform the two control models in RTPJ (**Cond vs. Ling:** W=219, p=.49; **Cond vs. Narr:** W=265, p=.16; **Emo vs. Ling:** W=247, p=.27, **Emo vs. Narr:** W=285, p=.07); there was additionally no difference in the fit of the condition and emotion models to the RTPJ RDM (W=195, p=.69).

Across all ROIs, the condition and emotion models did not outperform the control models. There were significant model-by-ROI interactions such that the condition model fit the neural data significantly better than the linguistic and narrative control models in LTPJ, relative to the RTPJ The condition model did not outperform the emotion model across all ROIs, but there were again model-by-ROI interactions such that the condition model fit the neural data better than the emotion model in LTPJ and PC, relative to RTPJ (see Supplementary Table 2 for statistics and Supplementary Figure 7).

Experiment 2: Comparisons of Neurotypical Children and Children with an ASD Diagnosis
The condition model fit the RTPJ responses in neurotypical children marginally better than children diagnosed with ASD (M(SE) NT=.03(.01), ASD=.001(.01); effect of group: b=-.39, t=-1.9, p=.06, effect of motion: b=-.31, t=-3.3, p=.001, no group-by-motion interaction). The group effect was significant in a regression that additionally controlled for non-verbal IQ (effect of group: b=-.52, t=-2.4, p=.02, effect of IQ: b=.01, t=.10, p=.92, effect of motion: b=-.28, t=-2.9, p=.005); this group difference remained significant with a Bonferroni correction for multiple comparisons (two tests; α=.025). The fit of the emotion model to RTPJ responses did not differ across children with and without ASD (M(SE) NT=-.002(.01), ASD=.009(.01); effect of group: b=.15, t=-.97, p=.34).

Across all ToM ROIs, there was a marginal effect of group on the condition model fit (effect of group (ASD): b=-.42, t=-1.96, p=.053, effect of ROIs: bs<|.15|, ts<|1.1|, ps>.3, effect of motion: b=-.21, t=-1.9, p=.06). Additionally, there were significant ROI-by-motion interactions, such that the condition model fit was higher in RTPJ, relative to other ROIs, in children who moved less: LTPJ: b=.32, t=2.2, p=.03; MMPFC: b=.35, t=2.4, p=.02, and a significant group-by-ROI (PC)-by motion interaction: b=.56, t=2.0, p=.04; all other interactions were not significant.

There was no effect of ASD diagnosis on the model fit of the emotion model across ROIs (effect of group (ASD): b=-.01, t=-.08, p=.94, effect of ROIs: bs<|.12|, ts<|.90|, ps>.3, effect of motion: b=-.10, t=-1.1, p=.29, ROI-by-motion interaction (MMPFC): b=.32, t=2.5, p=.01; no other significant interactions).

Supplementary Table 1

SubID	Age	Gender	Hand	ToM	IQ	Coil	N Art TP	MTrans (Pre)	MTrans (Post)
NT_1	7.05	M	L	0.613	135	Adult 32	35	0.08	0.07
NT_2	7.17	M	R	0.919	122	5yr 32	23	0.09	0.08
NT_3	8.95	F	R	0.886	132	Adult 32	94	0.19	0.11
NT_4	10.79	F	R	0.788	106	Adult 32	5	0.04	0.04
NT_5	8.18	M	R	0.939	124	5yr 32	55	0.13	0.08
NT_6	11.07	M	R	0.861	113	5yr 32	117	0.22	0.15
NT_7	8.72	M	R	0.800	116	Adult 32	110	0.16	0.13
NT_8	7.00	F	R	0.568	124	Adult 32	8	0.07	0.06
NT_9	5.66	F	R	0.513	127	5 or 7yr 32	8	0.03	0.03
NT_10	5.32	M	R	0.789	126	5yr 32	51	0.34	0.14
NT_11	10.94	F	R	0.974	126	Adult 32	66	0.08	0.07
NT_12	6.84	M	R	0.784	NA	5yr 32	147	0.15	0.09
NT_13	9.70	M	R	0.921	NA	NA	66	0.16	0.12
NT_14	6.94	M	NA	0.553	84	5yr 32	37	0.08	0.06
NT_15	10.07	M	NA	0.892	110	5yr 32	137	0.23	0.13
NT_16	8.99	F	L	1.000	103	7yr 32	98	0.19	0.11
NT_17	10.88	M	R	1.000	140	7yr 32	102	0.21	0.12
NT_18	10.19	M	R	0.943	111	Adult 32	96	0.18	0.15
NT_19	6.05	M	R	1.000	99	5yr 32	134	0.19	0.12
NT_20	8.33	M	R	1.000	125	7yr 32	133	0.13	0.06
NT_21	7.12	M	R	0.361	91	7yr 32	118	0.22	0.12
NT_22	9.26	M	NA	0.949	NA	7yr 32	101	0.23	0.11
NT_23	9.60	M	NA	0.949	124	7yr 32	5	0.06	0.06
NT_24	6.61	M	Ambi-R	0.949	129	7yr 32	96	0.12	0.07
NT_25	10.90	F	NA	0.947	104	Adult 32	78	0.08	0.06
NT_26	7.30	M	R	0.769	NA	7yr 32	115	0.16	0.08
NT_27	7.41	F	NA	1.000	NA	5yr 32	109	0.22	0.09
NT_28	7.87	M	NA	1.000	114	7yr 32	108	0.31	0.17
NT_29	7.22	F	R	0.897	116	7yr 32	77	0.19	0.10
NT_30	10.97	M	R	0.974	129	Adult 32	129	0.28	0.13
NT_31	9.54	M	R	0.949	142	Adult 32	168	0.23	0.13
NT_32	12.42	M	R	1.000	113	7yr 32	46	0.09	0.08
NT_33	12.42	M	R	1.000	136	7yr 32	16	0.08	0.07
NT_34	10.42	M	R	0.974	108	Adult 32	42	0.07	0.05
NT_35	12.60	M	R	0.895	132	7yr 32	66	0.15	0.12
NT_36	11.44	M	R	1.000	119	Adult 32	13	0.07	0.07
NT_37	7.03	M	R	0.811	92	7yr 32	59	0.12	0.09
NT_38	9.86	M	NA	0.923	122	7yr 32	47	0.07	0.06
NT_39	10.96	M	R	1.000	106	Adult 32	10	0.05	0.05
NT_40	12.29	M	R	NA	108	Adult 32	8	0.05	0.05
NT_41	12.27	M	NA	1.000	101	Adult 32	90	0.24	0.15
NT_42	12.39	M	NA	0.943	113	7yr 32	83	0.16	0.14
NT_43	8.68	M	R	1.000	101	5yr 32	60	0.10	0.07
NT_44	7.01	M	R	0.872	132	7yr 32	107	0.11	0.06
NT_45	10.75	M	R	0.923	126	7yr 32	130	0.21	0.09
NT_46	8.93	M	R	0.872	126	7yr 32	102	0.14	0.06
NT_47	8.92	M	R	0.897	NA	7yr 32	94	0.11	0.09
NT_48	8.49	M	R	0.947	127	5yr 32	40	0.07	0.06
NT_49	9.76	M	R	0.947	128	7yr 32	35	0.13	0.11
NT_50	8.41	M	R	0.846	114	7yr 32	76	0.11	0.07
NT_51	9.15	M	R	0.872	105	5yr 32	101	0.09	0.05
NT_52	5.26	M	R	0.816	107	7yr 32	144	0.23	0.13
NT_53	8.41	F	R	0.914	124	5yr 32	21	0.02	0.02
NT_54	5.38	F	R	0.743	103	Adult 32	18	0.05	0.04
NT_55	7.53	M	R	0.895	110	7yr 32	132	0.23	0.08
NT_56	7.38	F	R	0.811	118	5yr 32	15	0.07	0.06
NT_57	7.02	F	R	0.921	109	5yr 32	5	0.06	0.06
NT_58	8.68	M	R	0.974	96	7yr 32	163	0.30	0.16
NT_59	7.93	M	R	0.676	125	7yr 32	123	0.16	0.08
NT_60	9.83	M	R	0.846	128	7yr 32	104	0.16	0.06
NT_61	7.51	F	R	0.949	130	5yr 32	65	0.11	0.08
NT_62	6.12	M	R	0.757	112	Adult 32	38	0.14	0.11
NT_63	9.42	M	R	0.872	109	7yr 32	125	0.16	0.11
NT_64	9.20	F	R	0.941	132	Adult 32	136	0.16	0.11
NT_65	7.61	M	R	0.838	122	7yr 32	100	0.20	0.08
NT_66	10.12	M	R	0.921	124	7yr 32	122	0.13	0.08
NT_67	6.23	F	R	0.853	99	7yr 32	169	0.23	0.15
NT_68	5.84	M	R	0.641	115	5yr 32	19	0.08	0.06
NT_69	9.97	M	R	0.974	119	7yr 32	141	0.17	0.07
NT_70	7.73	M	R	0.897	115	7yr 32	116	0.16	0.10
NT_71	8.59	M	R	0.949	119	7yr 32	46	0.05	0.04
NT_72	8.09	M	L	0.889	105	Adult 32	211	0.23	0.09
NT_73	6.04	M	R	0.842	143	7yr 32	84	0.23	0.13
NT_74	6.03	M	R	0.949	117	7yr 32	151	0.22005	0.1207
NT_75	7.61	M	L	0.605	119	7yr 32	112	0.14306	0.092
NT_76	5.74	M	R	0.842	121	5yr 32	119	0.31011	0.17962
ASD_1	9.90	M	R	0.838	112	7yr 32	68	0.13	0.10
ASD_2	10.24	M	NA	0.897	NA	7yr 32	8	0.04	0.04
ASD_3	9.14	M	NA	0.737	NA	7yr 32	83	0.10	0.08
ASD_4	6.09	M	R	0.410	129	5yr 32	99	0.24	0.13
ASD_5	7.38	M	R	0.289	94	5yr 32	109	0.25	0.17
ASD_6	9.43	M	R	0.795	136	7yr 32	115	0.18	0.12
ASD_7	10.78	M	R	0.641	83	7yr 32	109	0.21	0.15
ASD_8	11.07	M	Ambi-R	0.821	99	Adult 32	73	0.30	0.20
ASD_9	12.05	M	R	0.946	95	7yr 32	41	0.13	0.11
ASD_10	8.60	M	R	0.351	80	7yr 32	52	0.08	0.05
ASD_11	9.99	M	L	0.795	96	Adult 32	112	0.17	0.12
ASD_12	NA	M	R	0.974	130	NA	15	0.04	0.04
ASD_13	8.74	M	R	0.400	100	7yr 32	130	0.18	0.12
ASD_14	5.61	M	R	0.618	100	5yr 32	96	0.15	0.11
ASD_15	10.22	M	R	0.949	129	7yr 32	42	0.09	0.08
ASD_16	8.84	M	R	0.833	126	7yr 32	160	0.22	0.12
ASD_17	9.97	F	R	0.556	116	7yr 32	55	0.14	0.11
ASD_18	10.38	M	NA	0.969	130	7yr 32	14	0.06	0.06
ASD_19	6.80	M	L	0.846	90	5yr 32	174	0.31	0.14
ASD_20	10.76	F	R	0.974	126	7yr 32	16	0.06	0.05
ASD_21	11.90	M	R	1.000	116	Adult 32	102	0.16	0.10
ASD_22	7.79	M	R	0.500	100	7yr 32	35	0.10	0.09
ASD_23	10.14	M	R	0.949	121	7yr 32	167	0.22	0.09
ASD_24	9.87	F	NA	0.974	137	7yr 32	4	0.03	0.03
ASD_25	7.69	M	L	0.718	82	7yr 32	66	0.11	0.07
ASD_26	8.56	M	R	0.718	101	7yr 32	44	0.15	0.11
ASD_27	10.98	M	R	0.949	126	7yr 32	59	0.14	0.13
ASD_28	12.87	M	R	0.947	126	7yr 32	52	0.09901	0.06953
ASD_29	9.05	F	R	0.872	95	Adult 32	50	0.11332	0.09519

Supplementary Table 1. Experiment 2 Participant Demographics. Experiment 2 participants included n=76 neurotypical children (NT) and n=29 children with Autism Spectrum Disorder (ASD), and therefore enabled us to test for developmental change in neural response patterns with age, and across samples with varying ToM reasoning abilities. SubID indicates sample (NT or ASD) and number. Age is in years. "Ambi-R" indicates ambidextrous with right hand preference. ToM is proportion correct on the behavioral ToM task. N Art TP refers to the number of artifact timepoints during the fMRI scan, and MTrans (Pre) and (Post) refer to the average amount of translation (movement in x,y.z directions) between timepoints during the scan, calculated before (Pre) or after (Post) artifact timepoint exclusion (see Methods in main text for more details).

Supplementary Table 2

Condition vs. Linguistic	Experiment 1	Experiment 2 (NT)	Experiment 2 (ASD)
Model (Ling)	**b=-.33, t=-3.5, p=.004**	**b=-.43, t=-5.7, p=2.6x10^{-8}**	b=.10, t=.41, p=.68
ROI (LTPJ)	b=.14, t=1.1, p=.27	b=.05, t=.45, p=.65	**b=.62, t=2.5, p=.01**
ROI (MMPFC)	b=.13, t=1.02, p=.31	b=.02, t=.18, p=.85	b=.23, t=.93, p=.36
ROI (PC)	b=.29, t=2.3, p=.02	b=-.10, t=-.93, p=.35	b=.33, t=1.3, p=.19
Model (Ling) x ROI (LTPJ)			**b=-.90, t=-2.5, p=.01**
Model (Ling) x ROI (MMPFC)			b=-.30, t=-.84, p=.40
Model (Ling) x ROI (PC)			b=-.44, t=-1.2, p=.22
Condition vs. Narrative	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Narr)	**b=-.63, t=-7.2, p=2.6x10^{-12}**	**b=-.56, t=-3.6, p=.0003**	b=-.19, t=-.83, p=.41
ROI (LTPJ)	b=.15, t=1.2, p=.22	b=.11, t=.74, p=.46	**b=.74, t=3.2, p=.002**
ROI (MMPFC)	b=.17, t=1.4, p=.17	b=-.15, t=-.96, p=.34	b=.28, t=1.2, p=.24
ROI (PC)	b=.14, t=1.2, p=.24	b=-.14, t=-.92, p=.36	b=.39, t=1.7, p=.09
Model (Narr) x ROI (LTPJ)		b=.07, t=.30, p=.77	**b=-.78, t=-2.4, p=.02**
Model (Narr) x ROI (MMPFC)		**b=.47, t=2.2, p=.03**	b=-.45, t=-1.4, p=.17
Model (Narr) x ROI (PC)		**b=.47, t=2.1, p=.03**	b=-.28, t=-.85, p=.39
Emotion vs. Linguistic	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Ling)	b=-.13, t=-1.5, p=.13	b=-.04, t=-.58, p=.56	b=.02, t=.17, p=.87
ROI (LTPJ)	b=.12, t=.94, p=.35	b=-.04, t=-.34, p=.74	b=-.22, t=-1.2, p=.24
ROI (MMPFC)	b=-.07, t=-.60, p=.55	b=.16, t=1.4, p=.15	b=-.03, t=-.18, p=.86
ROI (PC)	b=.14, t=1.1, p=.27	b=-.02, t=-.23, p=.82	b=-.27, t=-1.4, p=.15
Emotion vs. Narrative	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Narr)	**b=-.40, t=-4.5, p=9.9x10^{-6}**	b=.08, t=1.1, p=.28	b=-.19, t=-1.5, p=.14
ROI (LTPJ)	b=.12, t=.96, p=.34	b=.07, t=.60, p=.55	b=-.12, t=-.67, p=.50
ROI (MMPFC)	b=-.07, t=-.52, p=.61	**b=.23, t=2.1, p=.04**	b=-.09, t=-.51, p=.61
ROI (PC)	b=-.02, t=-.18, p=.85	b=.18, t=1.6, p=.11	b=-.20, t=-1.1, p=.27
Condition vs. Emotion	**Experiment 1**	**Experiment 2 (NT)**	**Experiment 2 (ASD)**
Model (Emo)	**b=-.18, t=-2.0, p=.04**	**b=-.38, t=-4.95, p=9.9x10^{-7}**	b=.12, t=.47, p=.64
ROI (LTPJ)	b=.07, t=.53, p=.60	b=.03, t=.27, p=.78	**b=.62, t=2.5, p=.01**
ROI (MMPFC)	b=-.03, t=-.20, p=.84	b=-.02, t=-.17, p=.87	b=.23, t=.95, p=.34
ROI (PC)	b=.13, t=1.03, p=.31	b=-.06, t=-.58, p=.56	b=.33, t=1.3, p=.18
Model (Emo) x ROI (LTPJ)			**b=-.79, t=-2.3, p=.03**
Model (Emo) x ROI (MMPFC)			b=-.24, t=-.68, p=.50
Model (Emo) x ROI (PC)			**b=-.76, t=-2.2, p=.03**

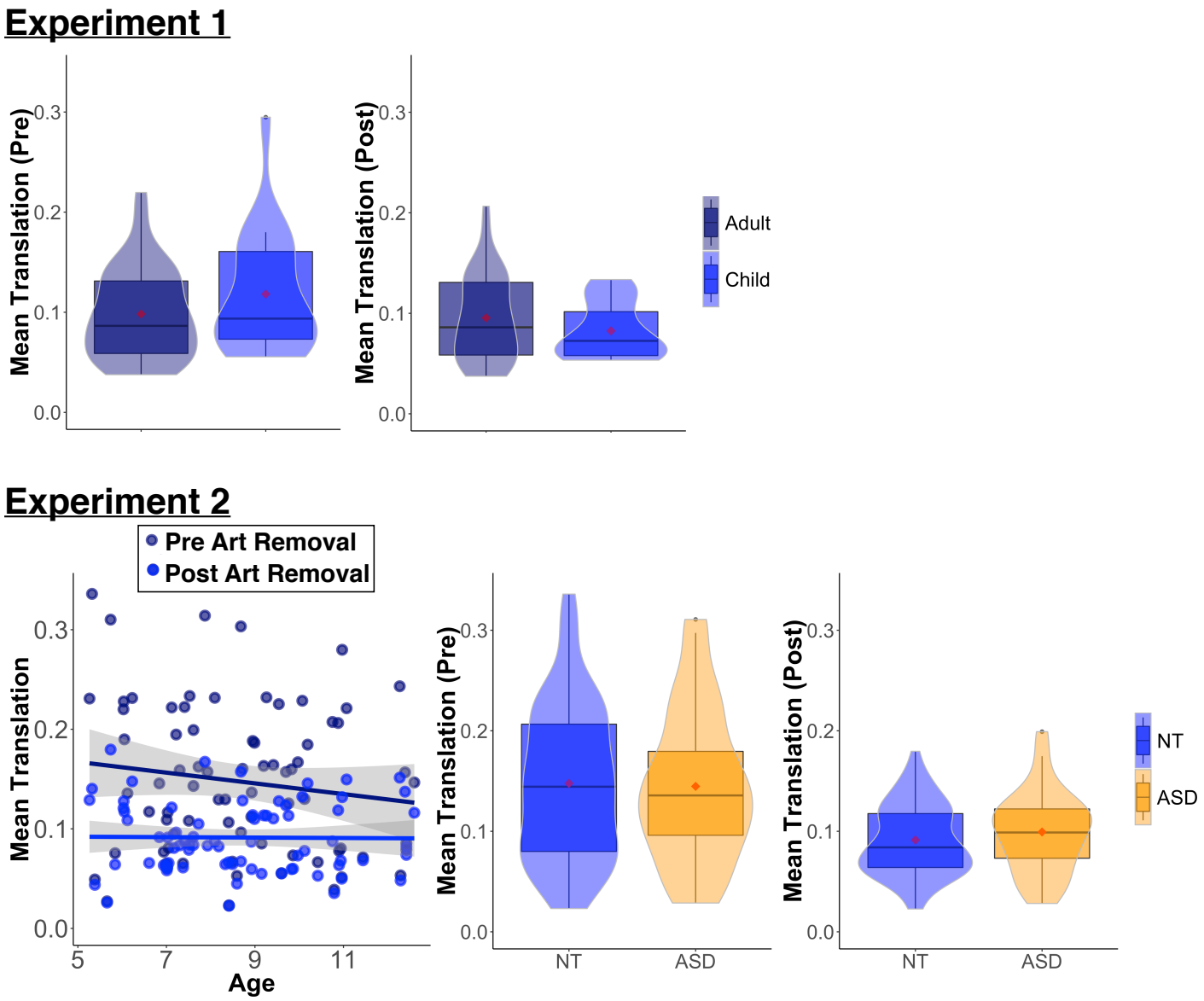
Supplementary Table 2. Statistical Results for Direct Comparisons of Model Fits Using Pearson Correlation. Full statistics (standardized beta values, t-values, and p-values) for linear mixed-effects regressions comparing the model fit of the planned ToM-relevant models (Condition, Emotion) to the control models (Linguistic, Narrative), and comparing the two ToM-relevant models to each other. Regressions tested for an effect of model (e.g., Condition vs. Linguistic) on the Kendall tau correlation values, which indicate fit to neural RDMs, and included region of interest (ROI) as a covariate. The right temporoparietal junction (RTPJ) was the reference ROI. Regressions also tested for significant Model-by-ROI interactions; non-significant interaction terms were removed from regressions (greyed cells). Significant results at a $p<.05$ threshold are shown in bold text. See Supplementary Figure 7 for visualization of results.

Supplementary Table 3

Condition Model: Euclidean Distance	[x,y,z] mm	$p_{cluster}$	$k_{cluster}$	p_{combo}	w_{combo}	$p_{voxel\ (FWE\text{-}corr)}$	Pseudo-t
Precuneus	**[10 -50 36]**	**0.0002**	**1129**	**0.0002**	**9.52**	**0.0002**	**6.31**
	[-4 -54 34]					0.0006	5.55
	[6 -62 32]					0.0034	5.23
Left Temporoparietal Junction	**[-54 -56 26]**	**0.0008**	**564**	**0.0002**	**9.52**	**0.0020**	**5.33**
	[-58 -48 20]					0.0078	5.06
	[-50 -44 26]					0.0928	4.44
Medial Prefrontal Cortex	**[-2 56 16]**	**0.0002**	**804**	**0.0002**	**9.52**	**0.0020**	**5.30**
	[6 48 18]					0.0446	4.64
	[-8 52 22]					0.0724	4.51
Right Middle Superior Temporal Sulcus	**[56 -22 -14]**	**0.1096**	**110**	**0.0376**	**4.60**	**0.0154**	**4.91**
Right Temporoparietal Junction	**[46 -50 22]**	**0.0016**	**491**	**0.0026**	**7.44**	**0.0196**	**4.86**
	[54 -58 14]					0.0668	4.54
	[58 -54 26]					0.2230	4.18
Left Inferior Parietal Cortex	**[-38 -60 46]**	**0.0100**	**285**	**0.0278**	**5.00**	**0.4144**	**3.94**
	[-54 -40 46]					0.6202	3.73
	[-34 4-8 40]					0.9748	3.15
Condition Model: Pearson Correlation Distance	**[x,y,z] mm**	$p_{cluster}$	$k_{cluster}$	p_{combo}	w_{combo}	$p_{voxel\ (FWE\text{-}corr)}$	**Pseudo-t**
Left Temporoparietal Junction	**[-48 8 -26]**	**0.0002**	**1848**	**0.0002**	**9.52**	**0.0002**	**6.43**
	[-48 -50 22]					0.0004	5.84
	[-58 -42 22]					0.0010	5.71
Precuneus	**[-2 -44 32]**	**0.0002**	**1551**	**0.0002**	**9.52**	**0.0002**	**6.35**
	[8 -44 28]					0.0002	6.27
	[10 -54 24]					0.0002	6.19
Medial Prefrontal Cortex	**[-8 50 24]**	**0.0002**	**753**	**0.0002**	**9.52**	**0.0010**	**5.71**
	[-20 52 26]					0.2028	4.39
	[-10 50 2]					0.2436	4.32
Right Inferior Frontal Gyrus	**[48 32 4]**	**0.0010**	**335**	**0.0006**	**8.82**	**0.0026**	**5.37**
	[50 12 2]					0.0378	4.85
	[42 40 4]					0.1804	4.42
Right Inferior Parietal Sulcus	**[46 -50 44]**	**0.0002**	**1123**	**0.0002**	**9.52**	**0.0046**	**5.28**
	[54 -56 12]					0.0094	5.14
	[56 -50 18]					0.0144	5.04
Right Anterior Temporal Sulcus	**[52 -4 -28]**	**0.0094**	**178**	**0.0094**	**6.05**	**0.0192**	**4.98**
	[44 2 -34]					0.0640	4.71
	[42 12 -30]					0.4302	4.10
Emotion Model: Pearson Correlation Distance	**[x,y,z] mm**	$p_{cluster}$	$k_{cluster}$	p_{combo}	w_{combo}	$p_{voxel\ (FWE\text{-}corr)}$	**Pseudo-t**
Left Middle Superior Temporal Sulcus	**[-64 -34 4]**	**0.0024**	**332**	**0.0018**	**7.91**	**0.0052**	**5.35**
	[-54 -28 0]					0.0084	5.23
	[-60 -22 -4]					0.3064	4.25
Right Middle Superior Temporal Sulcus	**[56 -28 -2]**	**0.0036**	**263**	**0.0044**	**6.95**	**0.0628**	**4.73**
	[46 -24 -4]					0.1754	4.45
	[60 -20 -6]					0.2626	4.31

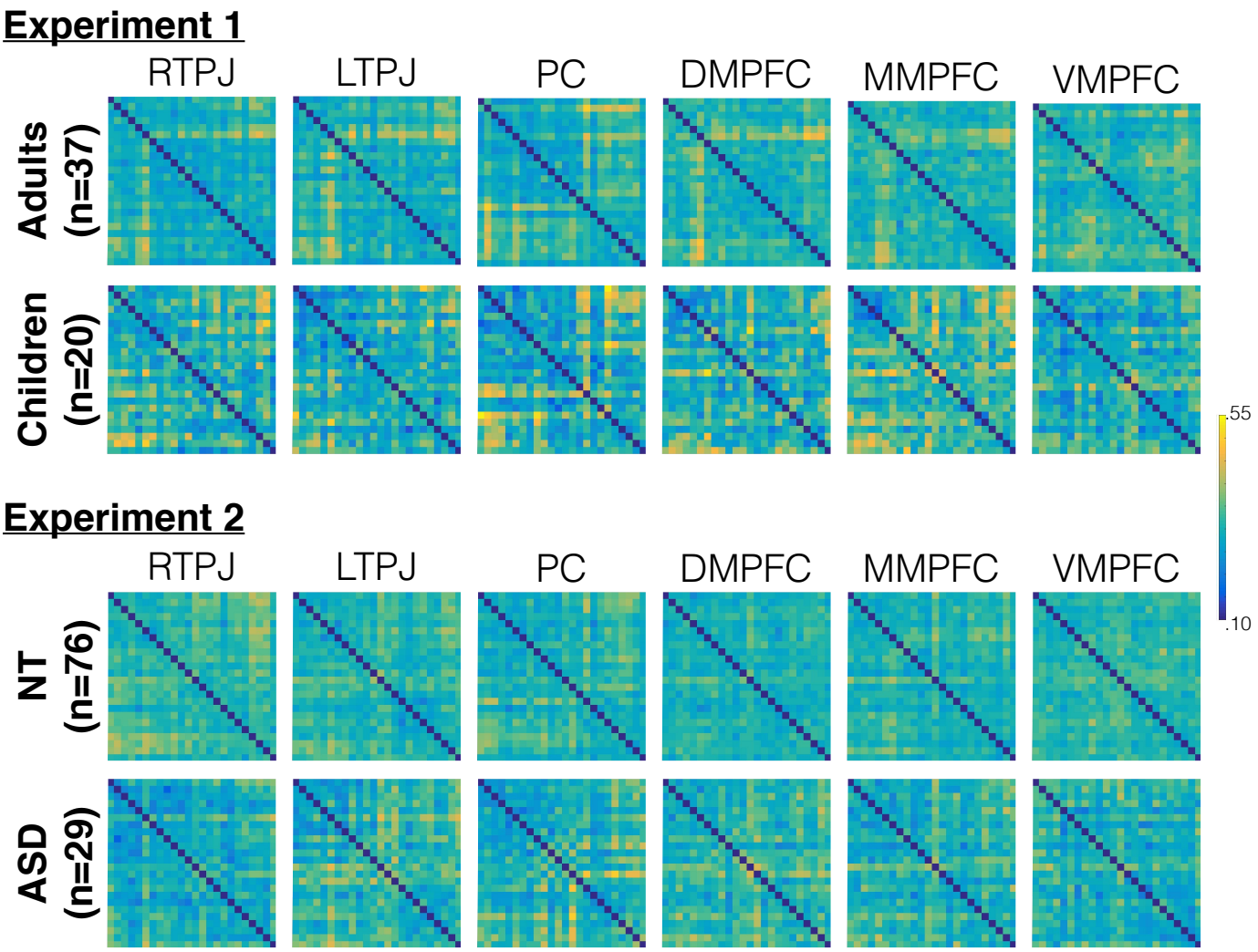
Supplementary Table 3. Searchlight Analysis for Condition and Emotion Model Fits. We conducted a searchlight analysis in the combined neurotypical child sample (n=96, across Exp. 1 and Exp. 2) to complement the ROI analyses, and to ensure that unpredicted effects did not go unnoticed. Detailed methods and a visualization of the results for the condition model (Euclidean distance) are provided in the main text. This table provides statistics and location information for clusters that survived correction for multiple comparisons (SnPM, $p<.05$), using Euclidean distance and Pearson correlation distance metrics (see Supplementary Figure 11 for visualization of clusters correlated with the emotion model at uncorrected thresholds (using Euclidean distance)).

Supplementary Figure 1



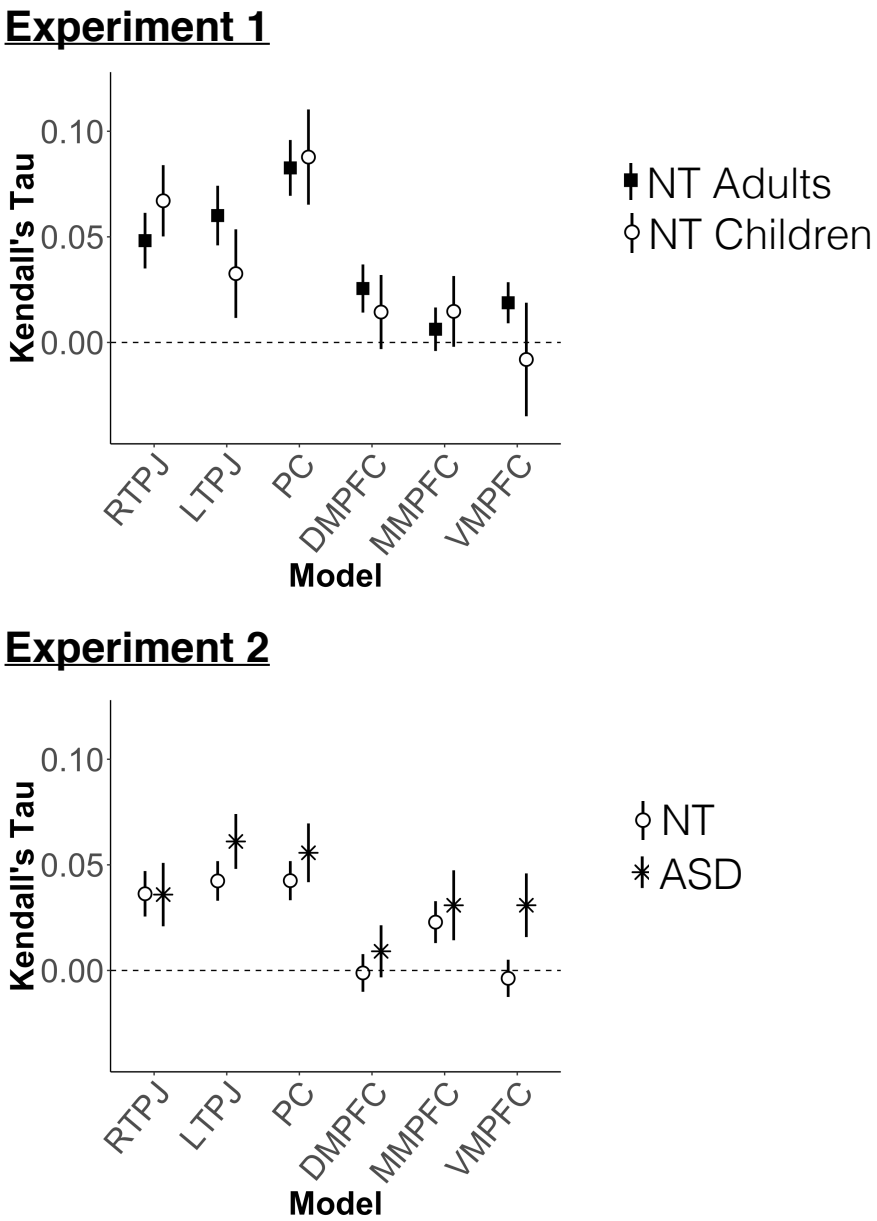
Supplementary Figure 1. Participant Motion. Experiment 1: Box plots show mean translation pre-artifact removal (left) and post-artifact removal (right) in adults (n=37, navy) and children (n=20, blue); violin outline indicates distribution. Red dots show group average. **Experiment 2:** Scatterplot (left) shows mean translation pre-artifact removal (navy) and post-artifact removal (blue) in neurotypical children (n=76) by age (in years, x-axis). Box plots show mean translation pre-artifact removal (middle) and post-artifact removal (right) in neurotypical children (blue) and in children diagnosed with Autism Spectrum Disorder (ASD, n=29, orange); violin outline indicates distribution. Red dots show group average. Mean translation refers to movement in x, y, z directions, in mm, between each timepoint during the fMRI scan (i.e., it is a measure of frame wise displacement).

Supplementary Figure 2



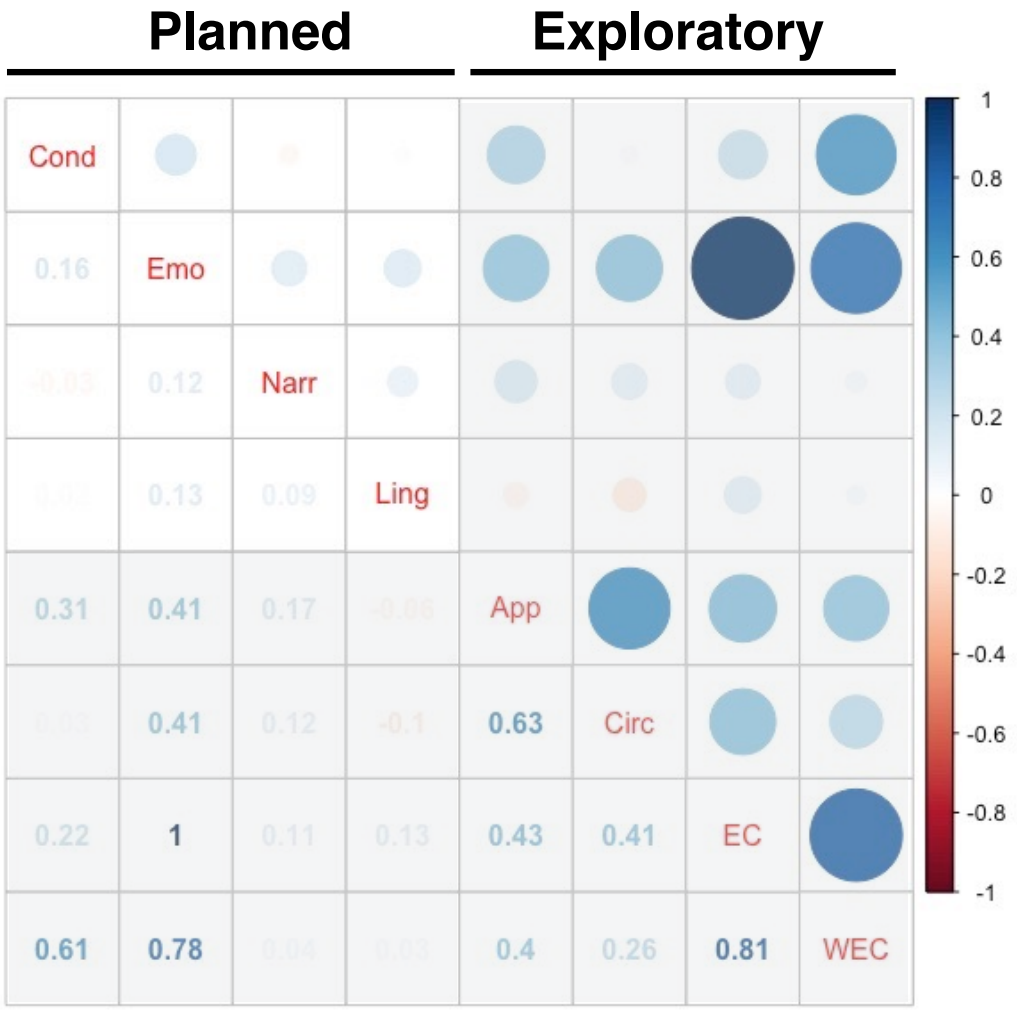
Supplementary Figure 2. Average Neural Representational Dissimilarity Matrices. Average neural RDMs per experiment, sample, and region of interest. Individual subject RDMs were created by extracting T-values from each voxel (n=80) within each ROI to each item, and calculating the Euclidean instance (square root of distance*distance) between each pair of stories, across voxels.

Supplementary Figure 3



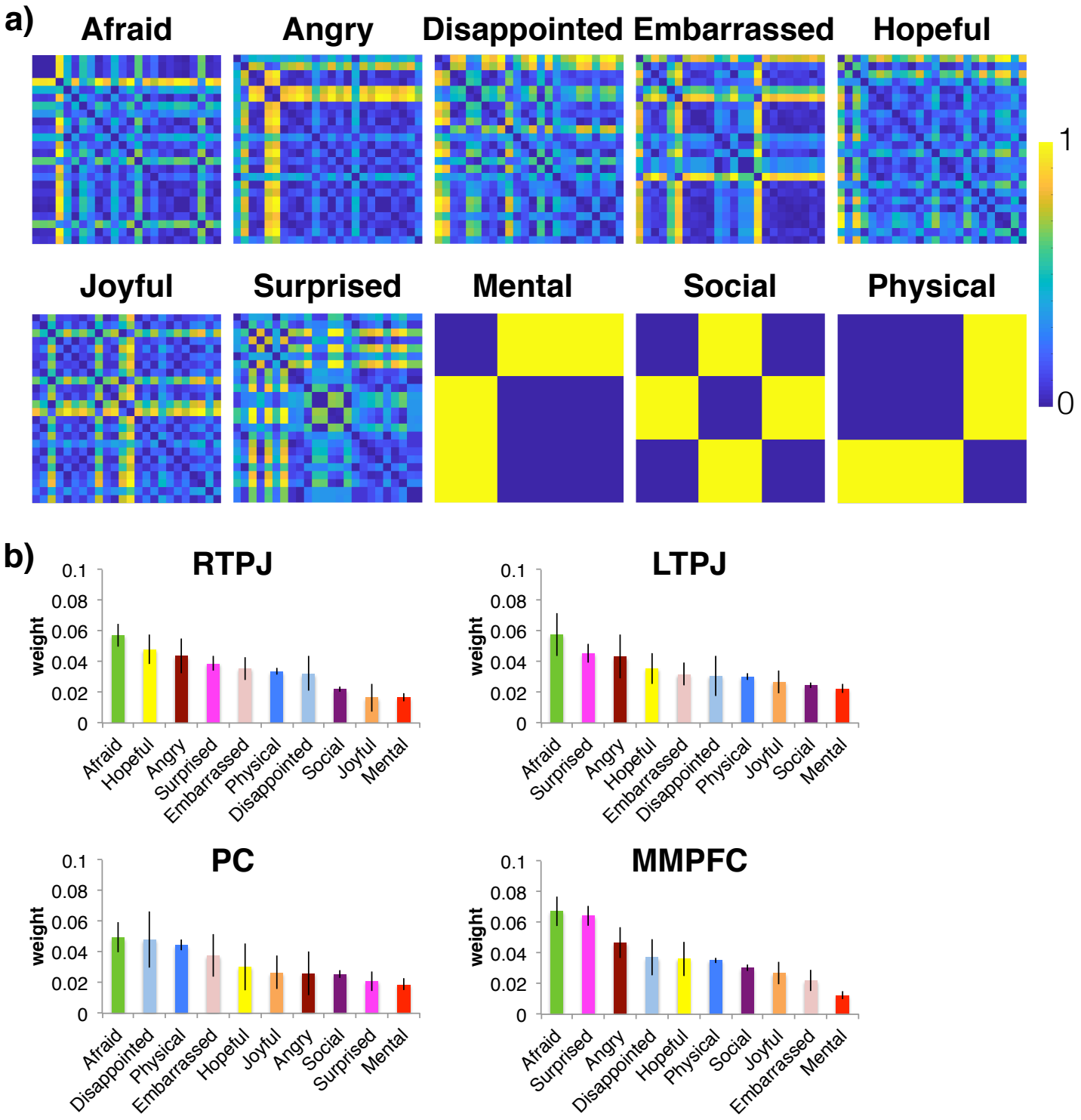
Supplementary Figure 3. Noise Ceilings per Experiment, Brain Region, and Sample. Noise ceilings were calculated as the Kendall's tau correlation (y-axis) between individual neural RDMs and the leave-one-child-out average neural RDM, per ROI. Point-line plots show mean correlations within each experiment and sample (Experiment 1: Adults (n=37, filled squares) and children (n=20, circles); Experiment 2: Neurotypical children (n=76, circles) and children diagnosed with Autism Spectrum Disorder (ASD; n=29, stars). Error bars reflect standard error of the mean.

Supplementary Figure 4



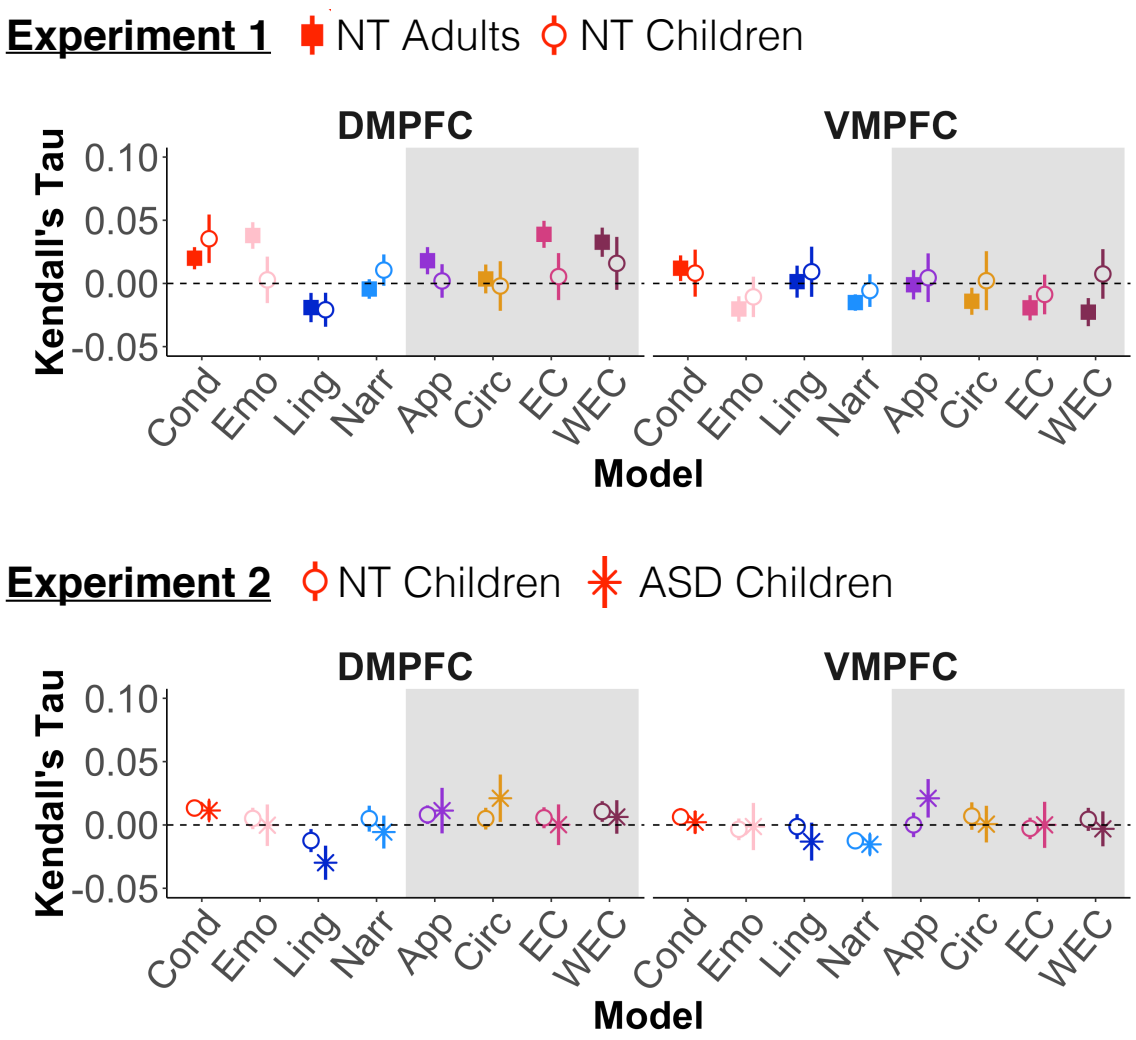
Supplementary Figure 4. Correlation Between Model RDMs. Correlation matrix shows the correlation between all model RDMs. Correlations values are indicated by r-values (bottom left) and visualized by size/shading of circles (top right). Planned Model RDMs were at most moderately positively correlated with one another (max correlation: r=.16, between Condition and Emotion models). The high correlations between Condition, Emotion, and EC (Emotion-Condition) and WEC (Weighted Emotion-Condition) models reflects shared features in these models.

Supplementary Figure 5



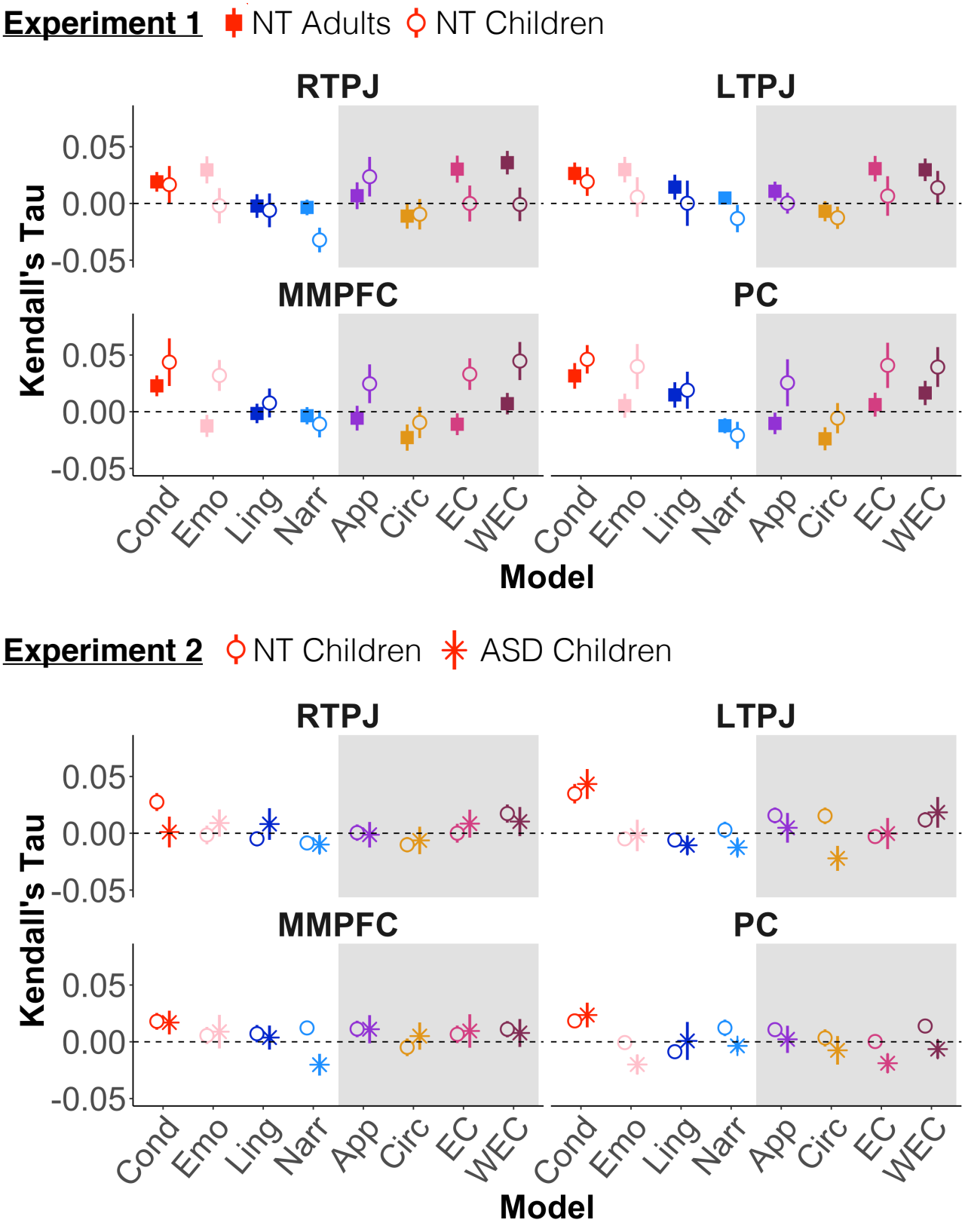
Supplementary Figure 5. Weighted Emotion-Condition Model. a) Model RDMs per emotion and condition features. 1 indicates maximal dissimilarity; 0 indicates similarity. **b)** Feature weights calculated using a non-negative least squares algorithm (Jozwik et al., 2016), predicting the average neural RDM, per ROI, in the Experiment 1 sample (n=57 children and adults). Features are ordered by feature weight magnitude, and color coded by feature.

Supplementary Figure 6



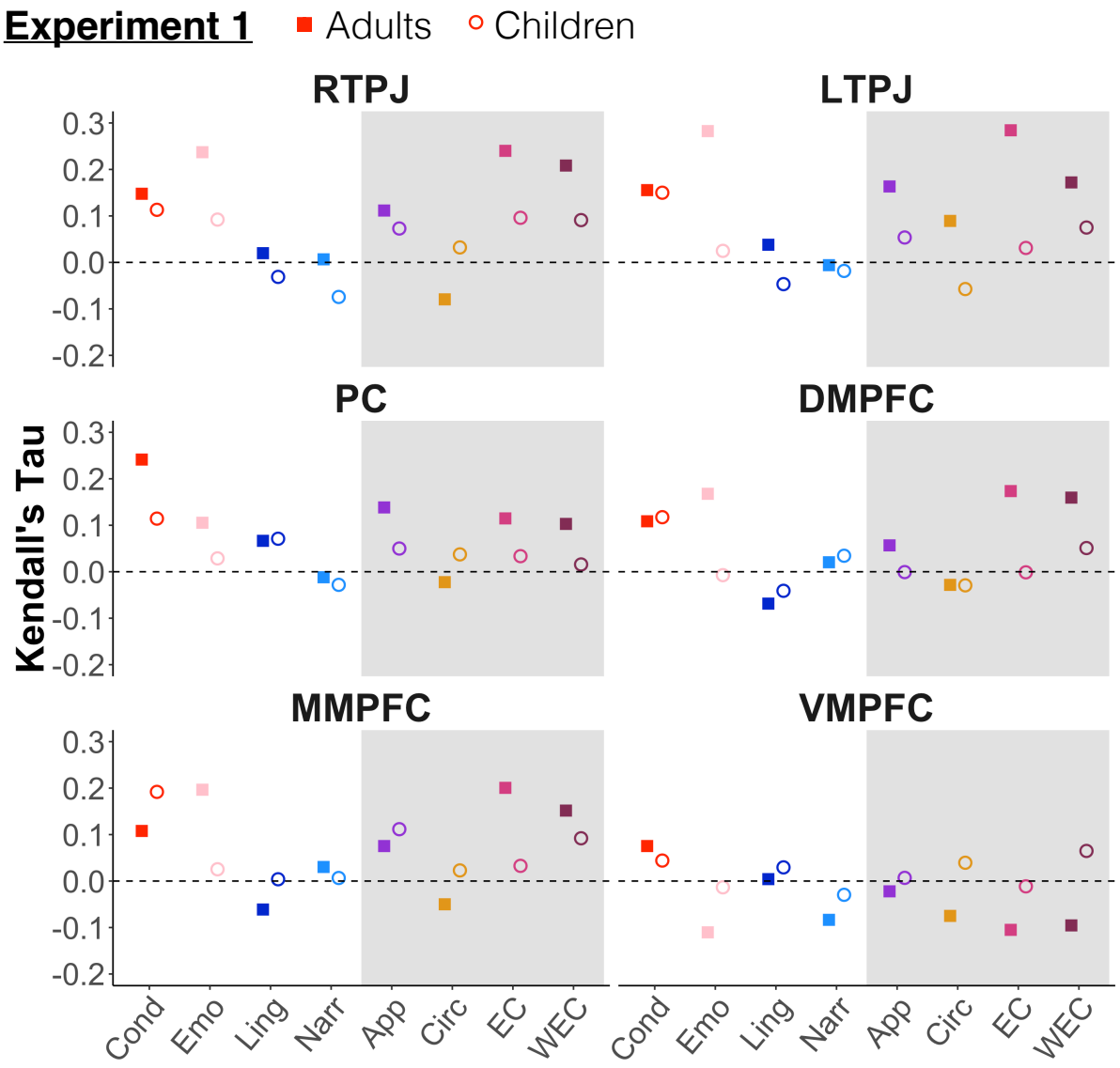
Supplementary Figure 6. Model Fits in DMPFC and VMPFC (Euclidean Distance). Plots show average model fits (Kendall tau correlation, y-axis) to individual neural RDMs extracted from DMPFC (left) and VMPFC (right) ROIs. These ROIs were excluded from all statistical analyses due to low noise ceilings (see Supplementary Figure 3); noise ceilings were inspected in both experiments prior to statistical analyses. ToM-relevant (Condition, Emotion) models are shown in red/pink; control (Linguistic, Narrative) models are shown in blues. The shaded area indicates exploratory models, which included a model based on abstract appraisal features (App, purple), a circumplex model based on valence and arousal (Circ, yellow), and models that included both emotion and condition features (EC, hot pink; W (weighted) EC, maroon)).

Supplementary Figure 7



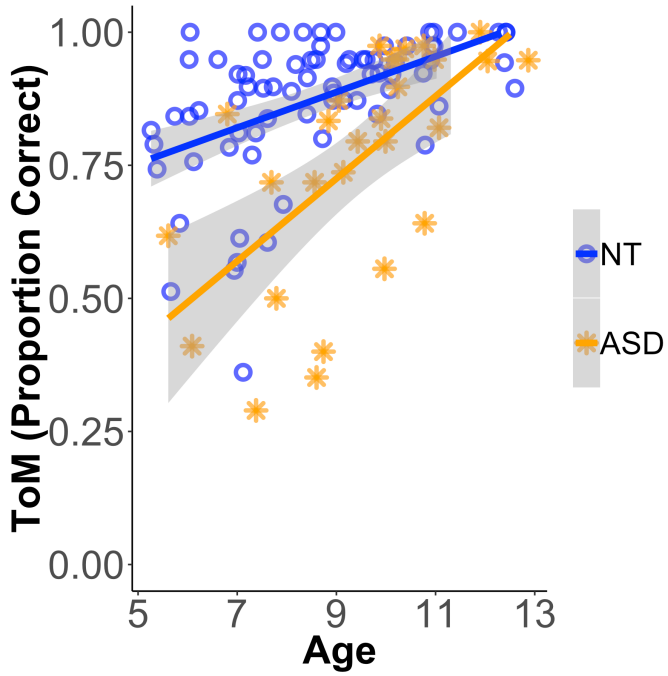
Supplementary Figure 7. Model Fits with Pearson Correlation Distance Dissimilarity Metric. Plots show average model fits (Kendall tau correlation, y-axis) to individual neural RDMs per ROI, model, and sample. See Supplementary Table 2 for statistics.

Supplementary Figure 8



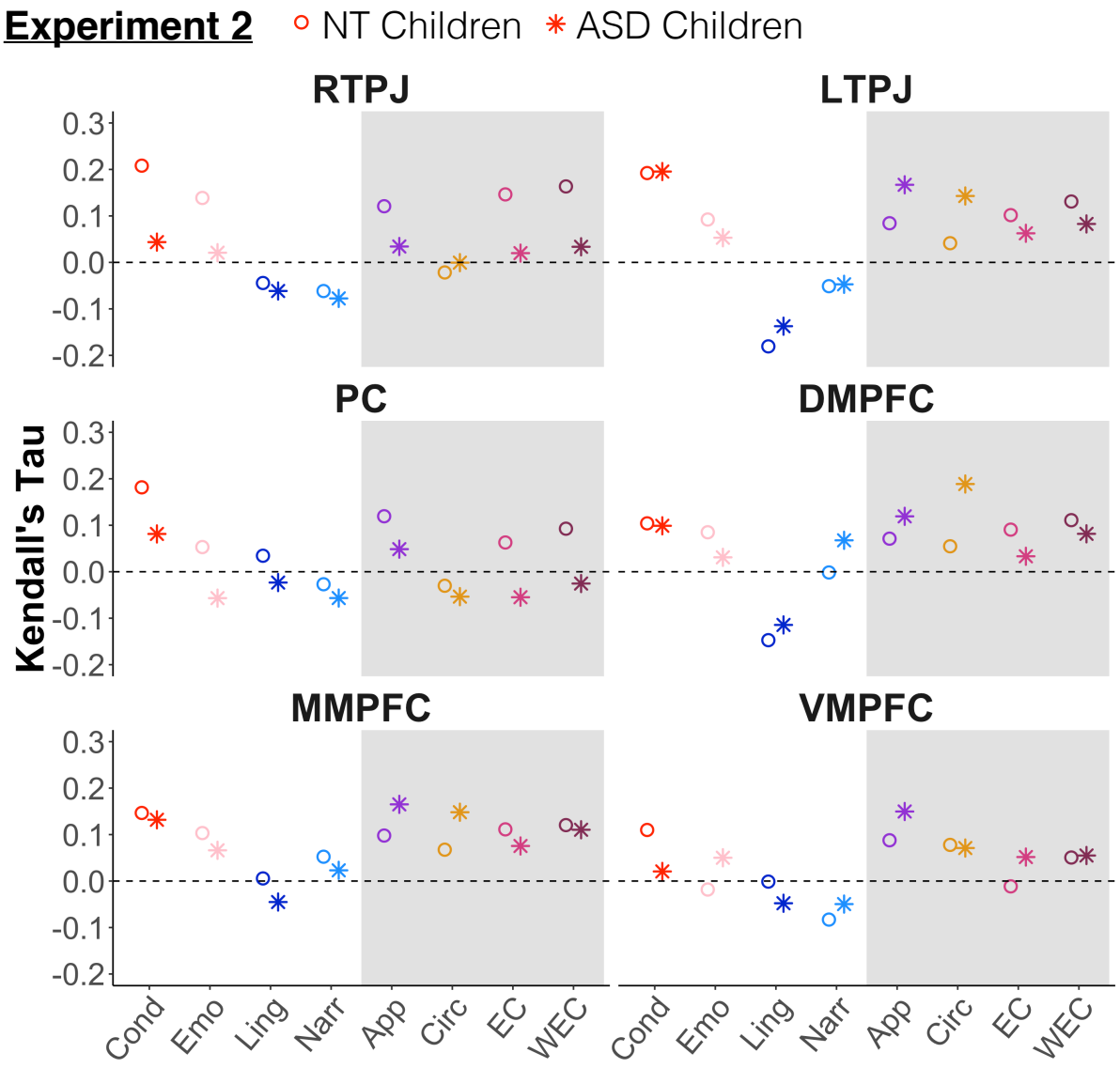
Supplementary Figure 8. Model Fits to Average Neural RDMs per ROI in Experiment 1. Point plots show the Kendall's Tau correlation between each model RDM and the average neural RDM per age group (n=37 adults, squares; n=20 children, circles), per ROI. Neural RDMs used Euclidean distance as the dissimilarity metric. ToM-relevant (Condition, Emotion) models are shown in red/pink; control (Linguistic, Narrative) models are shown in blues. The shaded area indicates exploratory models, which included a model based on abstract appraisal features (App, purple), a circumplex model based on valence and arousal (Circ, yellow), and models that included both emotion and condition features (EC hot pink; W (weighted) EC, maroon)).

Supplementary Figure 9



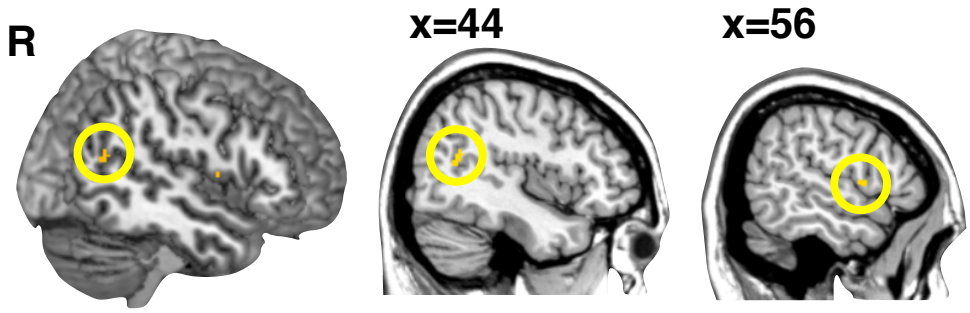
Supplementary Figure 9. Theory of Mind Behavior (Experiment 2). a) Performance on a ToM behavioral task (proportion of questions answered correctly; y-axis) by age (x-axis, years). Neurotypical children are shown in blue (n=75 (one child did not complete this task)); children diagnosed with Autism Spectrum Disorder (ASD) are shown in orange (n=29). Children with ASD performed worse on the ToM task than neurotypical children (b=-1.0, t=-5.6, $p=1.7x10^{-7}$), and ToM reasoning improved with age (cross-sectionally, in both samples and overall (b=.40, t=4.3, $p=4.3x10^{-5}$)). Age had a larger effect on performance in children with ASD, relative to neurotypical children (group-by-age interaction: b=.52, t=2.7, p=.008).

Supplementary Figure 10



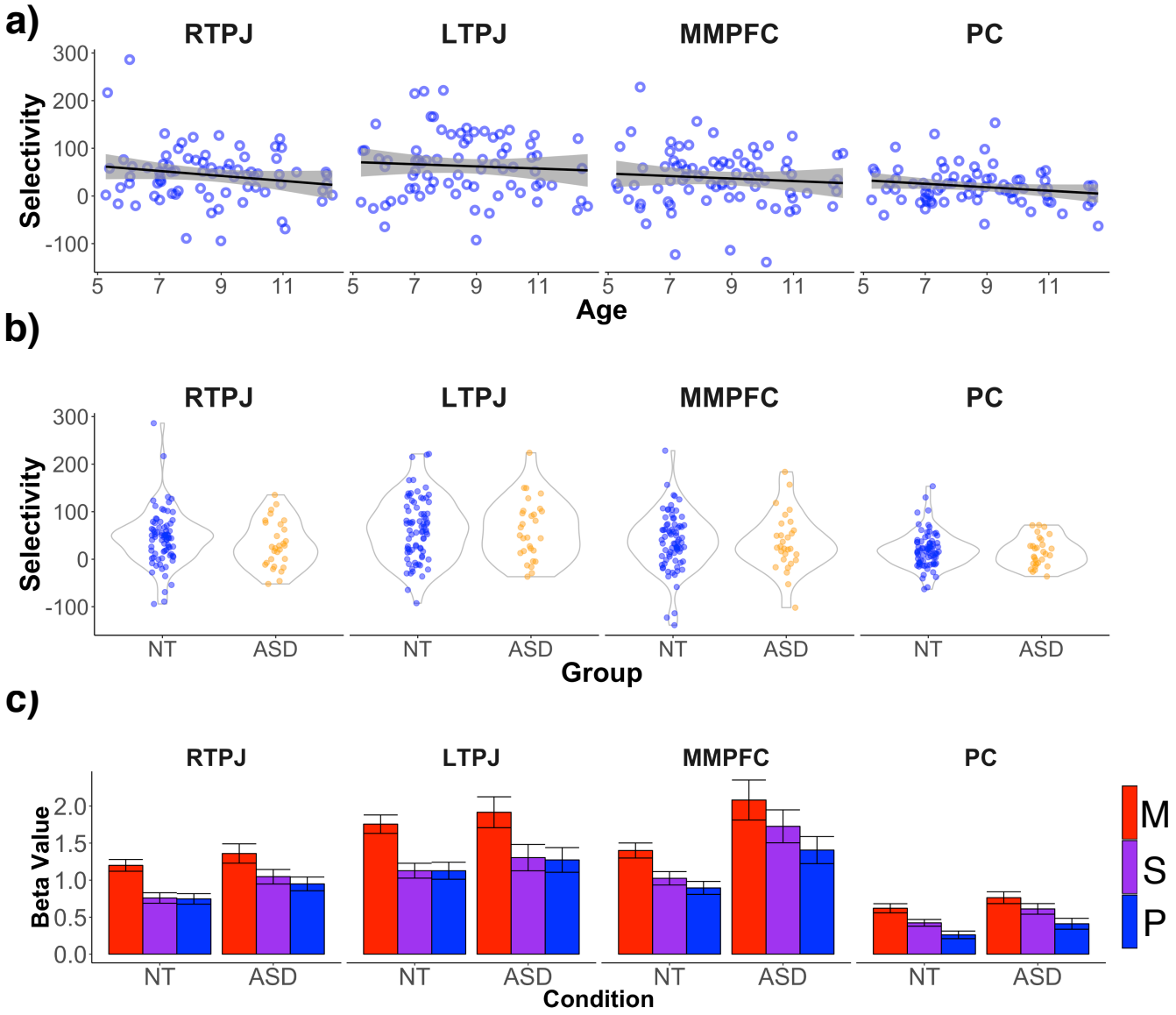
Supplementary Figure 10. Model Fits to Average Neural RDMs per ROI in Experiment 2. Point plots show Kendall's Tau correlation between each model RDM and the average neural RDM per group (n=76 neurotypical (NT) children, circles; n=29 children diagnosed with Autism Spectrum Disorder (ASD), stars), per ROI. RDMs used Euclidean distance to capture dissimilarity. ToM-relevant (Condition, Emotion) models are shown in red/pink; control (Linguistic, Narrative) models are shown in blues. The shaded area indicates exploratory models, which included a model based on abstract appraisal features (App, purple), a circumplex model based on valence and arousal (Circ, yellow), and models that included both emotion and condition features (EC, hot pink; W (weighted) EC), maroon).

Supplementary Figure 11



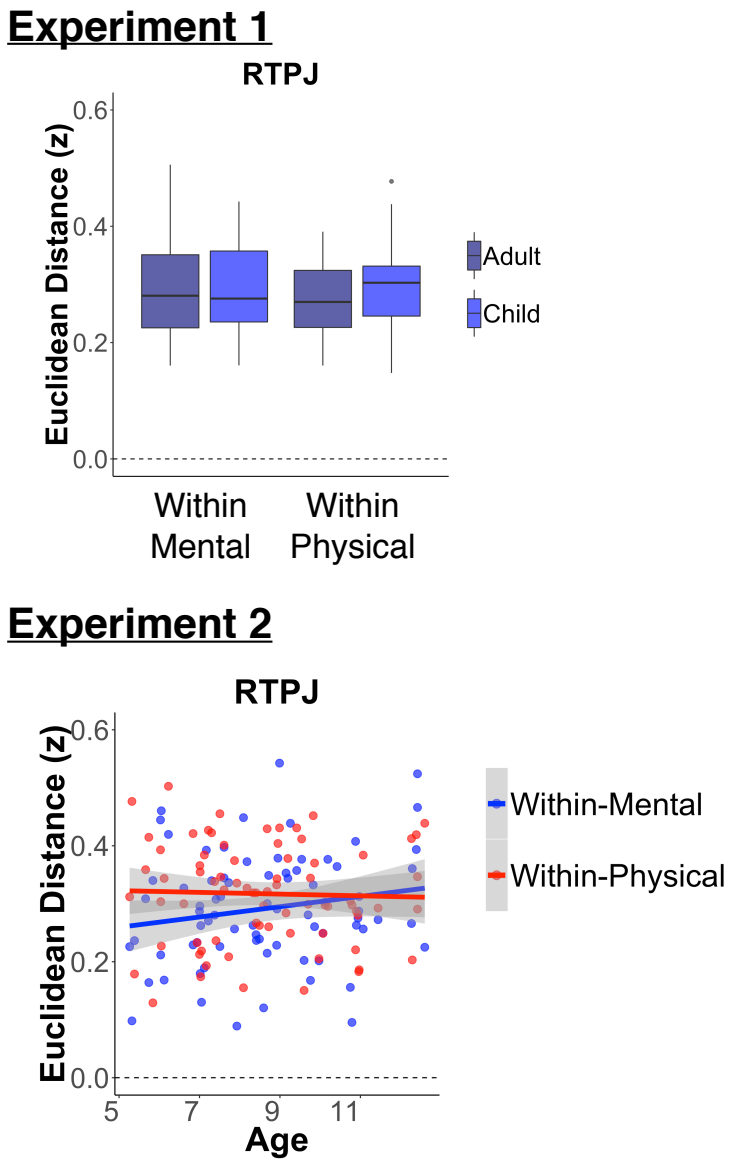
Supplementary Figure 11. Searchlight Analysis for Emotion Model Fit. We conducted a searchlight analysis in the combined neurotypical child sample (n=96, across Exp. 1 and Exp. 2) to complement the ROI analyses, and to ensure that unpredicted effects did not go unnoticed. Detailed methods and results of the condition model searchlight are described in the main text. In analyses that corrected for multiple comparisons ($p<.05$, SnPM), there were no significant voxels predicted by the emotion model. At more lenient thresholds ($p<.001$, $k=10$, uncorrected), response patterns in right superior temporal sulcus (rSTS) and premotor cortex correlated with the emotion model (peak voxel MNI coordinates (mm): rSTS: [44 -58 16], n voxels = 16, peak T = 3.46; premotor (two peaks): [58 8 6], [52 2 6], n voxels = 20, peak T = 3.50, 3.42). See Supplementary Table 3 for details of searchlight analysis results.

Supplementary Figure 12



Supplementary Figure 12. Univariate Responses per ROI and Group in Experiment 2. **a)** Scatterplots show selectivity index in neurotypical children (NT, n=76, blue) by age (years, x-axis) per region of interest. Bands show 95% confidence intervals. Selectivity did not increase significantly with age in neurotypical children. **b)** Violin plots show selectivity index per group (ASD, n=29, orange). Selectivity is calculated as average beta values for (Mental - Social)*100. Selectivity did not differ significantly between groups. **c)** Bar plots show the average beta value per condition (Mental (red), Social (purple), Physical (blue)), group, and region. Error bars show standard error from the mean.

Supplementary Figure 13



Supplementary Figure 13. Within-Condition Response Pattern Dissimilarity. Plots show the average normalized euclidean distance values (i.e., average response pattern dissimilarity) between stories within Mental and Physical conditions. Grey bands surrounding regression lines are 95% confidence intervals. We hypothesized that the response to individual Mental stories would become more distinct with age, such that pairwise dissimilarity between Mental stories would increase with age. While there was no evidence for change with age in the similarity of neural responses across Mental stories in Experiment 1 (i.e., between children (n=20, blue) and adults (n=37, navy); b=.10, t=.35, p=.73), there was a marginal increase in response dissimilarity across Mental stories with age among neurotypical children (n=76) in Experiment 2 (b=.20, t=1.7, p=.10).